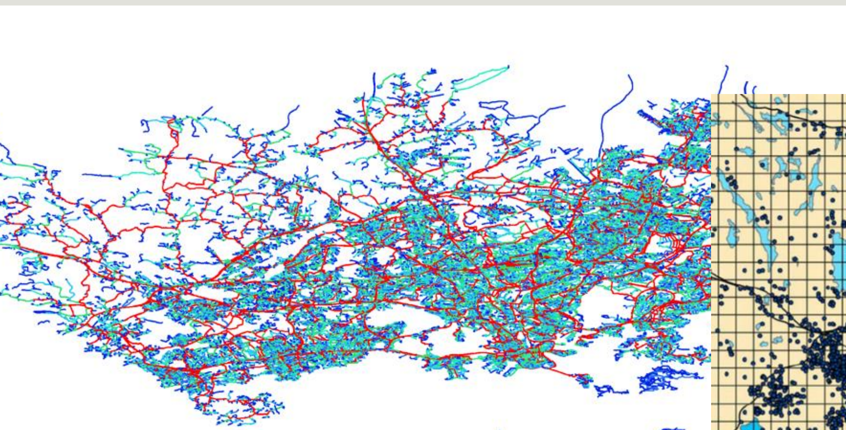
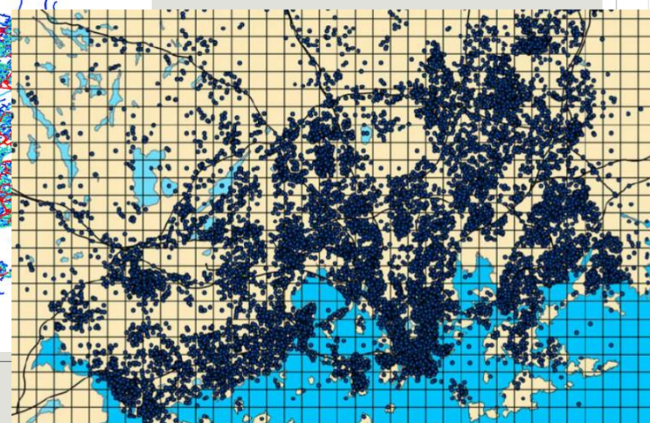
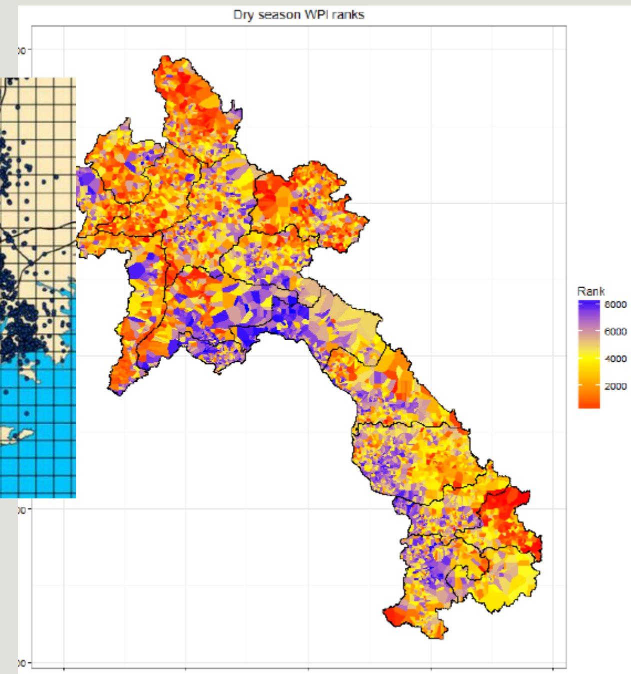
Advanced spatial analysis – Introduction, motivation, core concepts and the first data mining method







KIRSI VIRRANTAUS

GIS-E4020

10.1.2019

1.Goal and contents of the course

To learn more about selected advanced methods of geospatial analysis

To learn about theory of the methods

To achiceve skills in using the methods in ArcGIS, R and SPSS

To make a small project on using geospatial methods in problem solving

To become more interested in geospatial analysis by reading some research papers on the methods

The contents of the course

Spatial data mining methods

- Association rules, Clustering, Classification
- Geographically weighted regression
- Geographically weighted PCA
- Self organizing maps
- Trajectory and moving data mining

Agent based geosimulation

Fuzzy modeling of geoinformation

Classroom exercises and project work on SDM

date	topic	lecturer/teacher	exercises
10.1.2019 12.15... Kone1 201	Introduction Association rules for spatial data	Kirsi Virrantaus	
14.1.2019 10-12 U344		Jaakko Madetoja	1.Association rules, SPSS
17.1.2019 12.15... Kone1 201	Advanced spatial clustering methods	Marko Kallio	
21.1.2019 10-12 U344		Marko Kallio	2.Clustering, R
24.1.2019 12.15... Kone1 201	Geographically weighted regression Self organizing maps for spatial data	Jaakko Madetoja	
28.1.2019 10-12 U344		Jaakko Madetoja	3.GWR, ArcGIS
31.1.2019 12.15... Kone 1 201	Uncertainty in GWR Trajectory data mining	Jaakko Madetoja Kirsi Virrantaus	
7.2.2019 12.15... Kone 1 201	Agent based simulation with spatial data	Jussi Nikander	
14.2.2019 12.15... Kone 1 201	Fuzzy modeling Spatial decision tree	Vesa Niskanen Kirsi Virrantaus	
11.4.2019 13.00-16.00	EXAM		

Lectures and classroom exercises during Period III

Content of the first lecture

1) Intro to the course and recap on spatial analysis methods & introducing some new ones
2) Association Rules
and Spatial Co-location as an example on data mining

BREAK

3) Introduction to spatial data mining

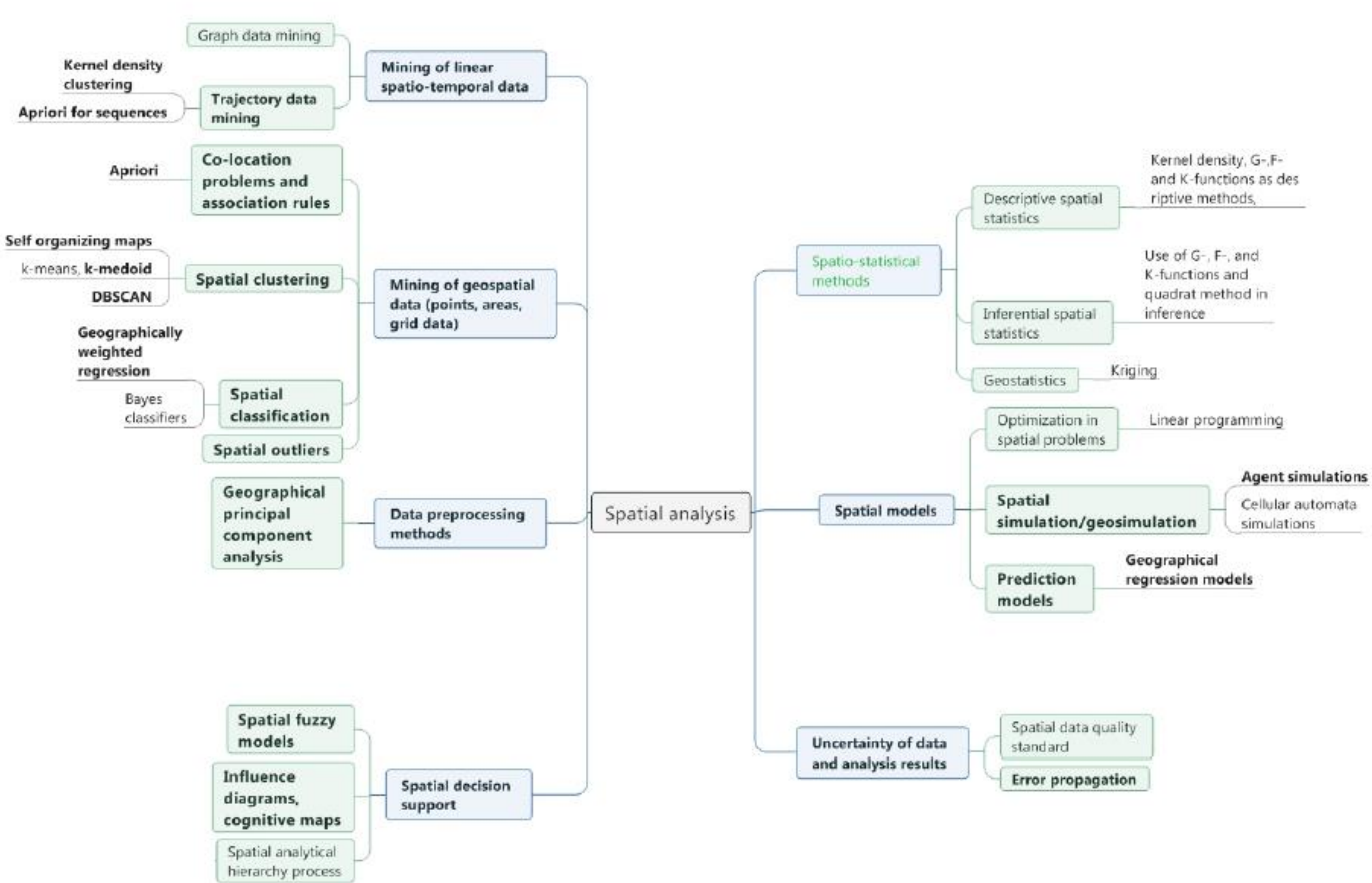
Learning goals of this first lecture
After the lecture:

You know what is the difference between spatio-statistical analysis and **geospatial data mining**

You can explain the differences between **data, information and knowledge**

You can explain the methods which can be applied in developing **non-spatial methods into spatial, so that the methods can manage spatial autocorrelation and spatial heterogeneity.**

You know **spatial association rules** –method


Spatial analysis
Mining of linear spatio-temporal data
Graph data mining
Trajectory data mining
Kernel density clustering
Apriori for sequences
Mining of geospatial data (points, areas, grid data)
Co-location problems and association rules
Apriori
Spatial clustering
Self organizing maps
k-means, k-medoid
DBSCAN
Spatial classification
Geographically weighted regression
Bayes classifiers
Spatial outliers
Data preprocessing methods
Geographical principal component analysis
Spatial decision support
Spatial fuzzy models
Influence diagrams, cognitive maps
Spatial analytical hierarchy process
Spatio-statistical methods
Descriptive spatial statistics
Kernel density, G-,F- and K-functions as des riptive methods,
Inferential spatial statistics
Use of G-, F-, and K-functions and quadrat method in inference
Geostatistics
Kriging
Spatial models
Optimization in spatial problems
Linear programming
Spatial simulation/geosimulation
Agent simulations
Cellular automata simulations
Prediction models
Geographical regression models
Uncertainty of data and analysis results
Spatial data quality standard
Error propagation

Review of your learning process
-make this exercise in 2-3 persons groups

List the methods that you have learned (which do not need any more teachning); in the group **mark with X** for each person who feels that s/he know the method already, **mark with XE** for each who has exercised the method and **K** if you have heard about the method

Clustering (list the methods like k-means, Dbscan, neural networks, SOM…what else?)

Classification (nearest neighbour classifier, Bayes classifiers, decision tree classifier… what else?)

Association rules

Fuzzy modeling

Uncertainty of analysis results

Agent simulation

2. Association pattern and association rule mining

The simplest example of association rule mining is the situation when we have a binary database (binary data matrix), with 0 and 1 values

Each column of the matrix (data base) represent a transaction

Transaction can be for example a set of items a customer buys in a shop

So-celled ”shopping basket”-example is the most popular example

Each row represents a transaction, transactions represents a shopping basket

Columns represent the item types bought

Frequent patterns mean an itemset that occurs in many transactions = in many shopping baskets

Shopping basket case

In a shopping basket
you have several items

The shopping basket is
a transaction
in which you have
things like
bread, wine, bananas…

All the items are in the
same basket at the same
time

Different customers have
different contents of the basket

It is interesting to see if some
items co-occur often in the
baskets

In the same way we can view
records/transactions
in a data base

Shopping baskets as a table of transactions

In this table there are 5 shopping baskets (rows)

Each column describe the contents of them

Baskets are called as transactions

Trans action	milk	brea d	butte r	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

2.1 Concepts used

Frequent itemsets = frequent patterns is the core concept

The relative frequency of a itemset in the data base is called as **support**

Diapers→Beer ; one of the most famous patterns in data mining (in the so-called ”shopping-basket” case)

Association rules describe models that occur often in the data base

Association rules are a result of analysis of transactions in a data base
- we try to search for associations between seemingly unrelated data in a relational database or other data repository

transaction = transaktio,
- ”tapahtuma” (tietojenkäsittelyssä)
- can be seen as a record of a file/relation in a table

Discovery of relationships within attributes of a relation is the simplest and most well-known data mining technique

Association rule

Is it the same than correlation ? But not exactly the same

A weaker form of correlation; no negative associations are found

In probabilistic terms association rule is the conditional probability: *P(Y|X); given that X has happened, the probability of Y*

In case the events X and Y are statistically independent P(Y|X) = P(Y)

Given that there is diapers in the shopping basket, the probability that there is also some beer.

Definitions

Following the original definition by Aggarwal the problem of association rule mining is defined as:

- Let $I = \{x_1, x_2, \ldots x_n\}$ be a set of n binary attributes called *items.*
- Let $D = \{t_1, t_2, \ldots t_n\}$ be a set of transactions called the *database*.
- Each transaction in *D* has a unique transaction ID and contains a subset of the items in *I*.
- A *rule* is defined as an *implication (seuraus)* of the form X→Y where X,Y belong to I and X∩Y = 0 (X ja Y ovat I:n osajoukkoja, joiden leikkaus =0)
- The sets of items (for short *itemsets*) *X* and *Y* are called *antecedent (=preceding event, edeltäjäosa)* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS, seuraajaosa) of the rule.

Support

Association rule is of the form *X→Y (c%);*

Association rule is characterized by two parameters: *support (suom. tuki)* and *confidence (suom. luottamus)*

***Support** is the relative frequency of an itemselt = a pattern in the rows of the matrix (in transactions), for example 40%*

The *support* supp of an itemset *C* is defined as the ***proportion of transactions in the data set which contain the itemset C**; support määritellään niiden transaktioiden osuudeksi datasetistä, jotka sisältävät alkiojoukon C (C is a subset of I)*

Minimum support is set by the users and is a parameter for the algorithm.

Confidence

If we have two itemsets X and Y, we can make an association rule X=>Y

The confidence of the rule is the fraction of transactions containing X that also contain Y

Confidence is calculated by dividing the support of XUY with the support of X

Confidence can be interpreted as *an estimate of the probability P(Y | X),* the **probability of finding the RHS** of the rule in transactions under the condition that these transactions **also contain the LHS**; *confidence tulkitaan estimaatiksi todennäköisyydestä että transaktioista löytyy Y ehdolla että niistä löytyy myös X*

Minimum confidence is a parameter of the algorithm set by the user.

2.2 Shopping baskets as a table

In this table there are 5 shopping baskets in each row

the columns of the table describe the contents of them

baskets are called as transactions

Transaction	milk	bread	butter	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Example on a rule and support and confidence of the rule

{butter, bread} => {milk }

- this example rule for the supermarket (shopping basket) could be meaning that if butter and bread is bought, customers also buy milk

- the itemset is {milk,bread,butter} and we have got the rule

{butter, bread} => {milk } (s=20%, c=50%)

- s=support means that 20% of transactions in the database contain {butter, bread, milk}
- c=confidence means that 50% of transactions with {butter, bread} also contain {milk }

Exercise 1

Can you find any other relevant rules

In the given data set?

What about milk and bread? Bread and milk? You can also think on 1-itemsets.

Trans action	milk	brea d	butte r	beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

2.3 Algorithms

Brute force algorithms are based on generating all candidate itemsets and count their support agaist the transaction database

This approach leads to a massive computation operation if the universe of items is large

The algorithms have been developed by several approaches, like so-called downward closure that helps to reduce the amount of possible candidate itemsets

The most popular algorithm of this type is called as **Apriori -algorithm**

Apriori – An algorithm to calculating frequent itemsets

First: all 1-itemsets (singletons) which are frequent (=exeed the given **support threshold**) are discovered

Second: all frequent itemsets are combined to form 2-itemsets; this set is parsed (jäsentää) to search for frequent 2-itemsets (exeeding the **given treshold)**

The process goes on: frequent 2-itemsets are combined to form 3-itemsets

Next step is to search for rules which satisfy **the minimum confidence requirement**

Given a frequent itemset {A,B,C}, all combinations are checked to see if they satisfy the confidence parameter c

Those that cross the threshold c are **association rules**

The attributes

ITEMS

Car CD Player	D
Car Alarm	A
TV	T
VCR	V
Computer	C

FREQUENT ITEMSETS

SUPPORT	ITEMSETS
100% (6)	A
83% (5)	C, AC
67% (4)	C, T, V, DA, DC, AT, AV, DAC
50% (3)	DV, TC, VC, DAV, DVC, ATC, AVC, DAVC

(Shekhar&Chawla)

Calculated supports for 1-,2- and 3-itemsets

Transactions

DATABASE

1	D A V C
2	A T C
3	D A V C
4	D A T C
5	D A T V C
6	A T V

ASSOCIATION RULES WITH CONFIDENCE = 100%

D → A (4/4)	D → A (4/4)	VC → A (3/3)
D → C (4/4)	D → A (3/3)	DV → A (3/3)
D → AC (4/4)	D → A (3/3)	VC → A (3/3)
T → C (4/4)	D → A (4/4)	DAV → A (3/3)
V → A (4/4)	D → A (3/3)	DVC → A (3/3)
C → A (5/5)	D → A (3/3)	AVC → A (3/3)

ASSOCIATION RULES WITH CONFIDENCE >=80%

C → D (4/5)	A → C (5/6)	C → DA (4/5)

2.4 How to present spatial associations

A simple approach in using association rule mining for spatial data is to apply spatial indexing or buffering.

Co-location is according to these spatial structures define the transactions.

In spatial data relationships can be described also by spatial predicates: equals, disjoint, touches, contains, covers, intersects, within, crosses, overlaps (Dimensionally Extended nine-Intersection Model (DE-9IM))

- Example: a country that is adjacent to the Mediterranean Sea is a wine-exporter (touches = adjacency)

Implementing spatial association rule mining: simple case

By using defined neighbourhood: grid or buffer.

- Association rules are generalized to data sets which **are indexed by space**
- Notion of transaction is replaced by **neighbourhood**
- In the following one exampe of our own works (from quite long time ago)

Application of Spatial Association Rules for Improvement of a Risk Model for Fire and Rescue Services

VĚRA KARASOVÁ, JUKKA MATTHIAS KRISP, KIRSI VIRRANTAUS

INSTITUTE OF CARTOGRAPHY AND GEOINFORMATICS

HELSINKI UNIVERSITY OF TECHNOLOGY

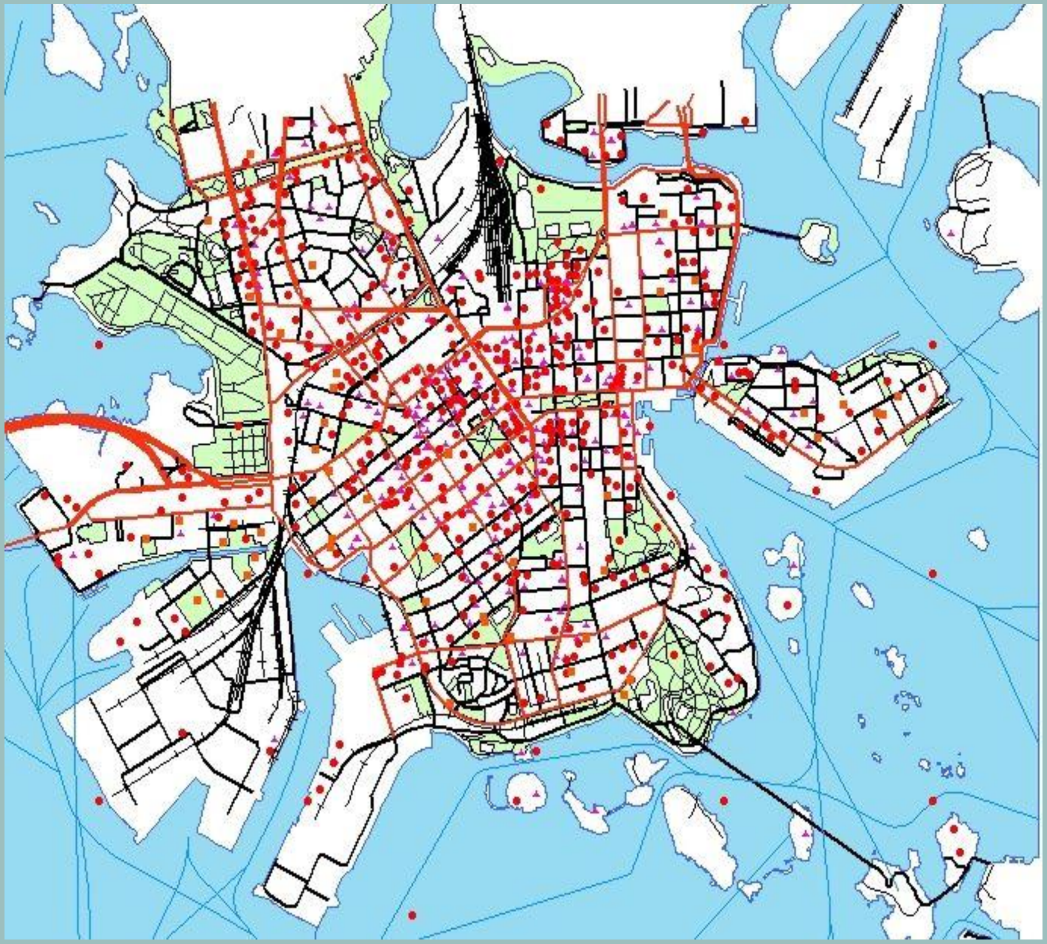
SCANGIS2005, STOCKHOLM

13TH-15TH JUNE 2005

Case study

- Register of Helsinki fire and rescue services
  - Incidents
- SeutuCD *[YTV, 2005]*
  - Kindergartens
  - Bars and restaurants
  - Main roads
  - Minor roads
  - Motorways
  - Paths
  - Railways
  - Waterways
  - Parks and cemeteries
  - Water

Study area: Helsinki city center



Method

Data pre-processing

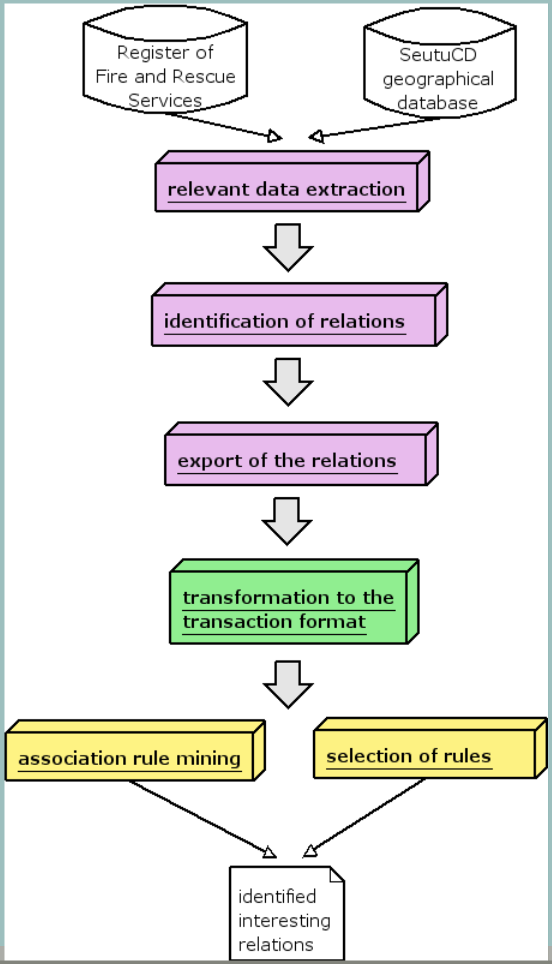
- extracton of relevant data
- definition of the neighbourhood

Transformation to the transaction format

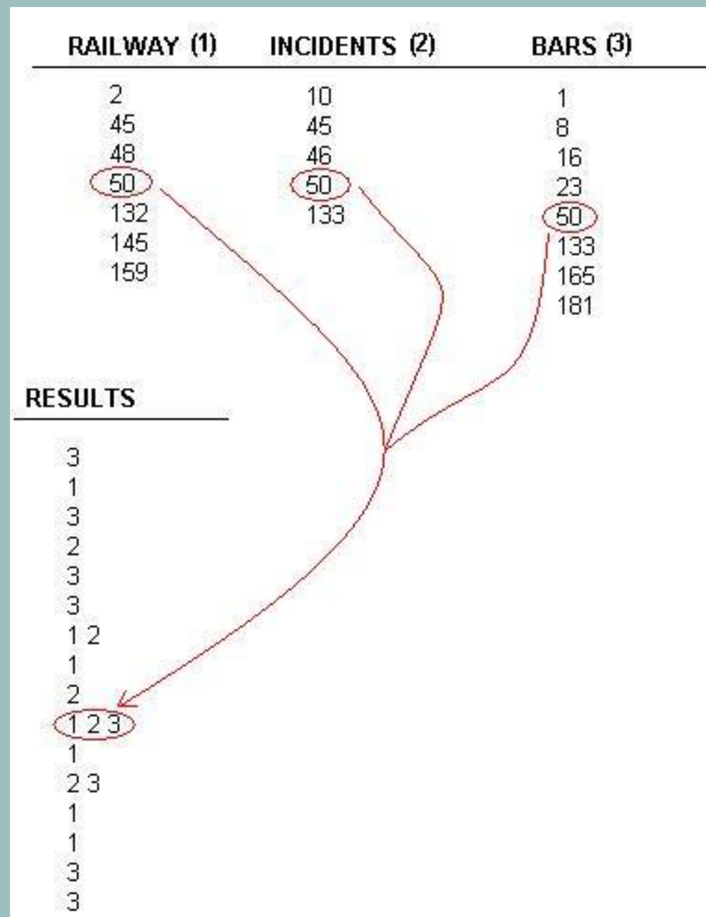
- integration

Association rule mining

- selection of only important rules (strong rules and application of syntactic constraints)



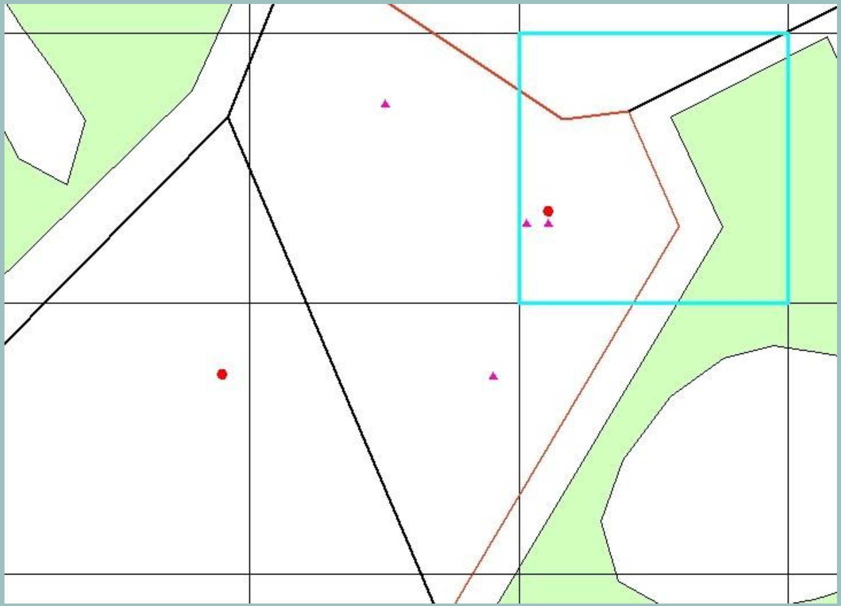
Transformation to transactions



Spatial indexing for co-location

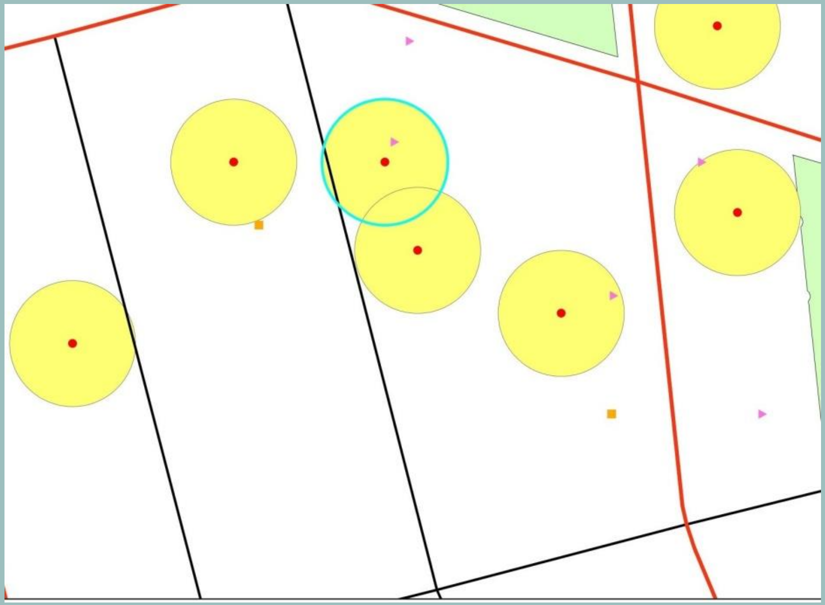
Grid approach

square regular grid over the whole study area (50 x 50 m)



Buffer approach

circular buffer around incidents (r = 50 m)



Association rule mining

Apriori algorithm implementation Gnome Data Mine tool
[Borgelt and Kruse, 2002] [Togaware, 2005]

Definition of constraints

- Minsupport = 0
- Minconfidence = 0
- Syntactic constraint: generate only rules with incidents

Results

bars and restaurants => incidents (1.7%; 40%)

incidents => main roads (2.2%; 30,4%)

incidents => minor roads (1.7%; 24.1%)

motorway => incidents (0%; 2.9%)

incidents => water (0.4%; 5.7%)

Conclusion

Definition of spatial data mining

Test the use on real data

Utilization of an existing tool originaly implemented for data mining

Useful method for exploring big amounts of data

Detection of implicit relations among selected objects

Possible use for identifying variables to improve the existing model for Fire and Rescue Services

Spatial predicates in association rule mining

Association rules can also be created by using spatial predicates, spatial relationships.

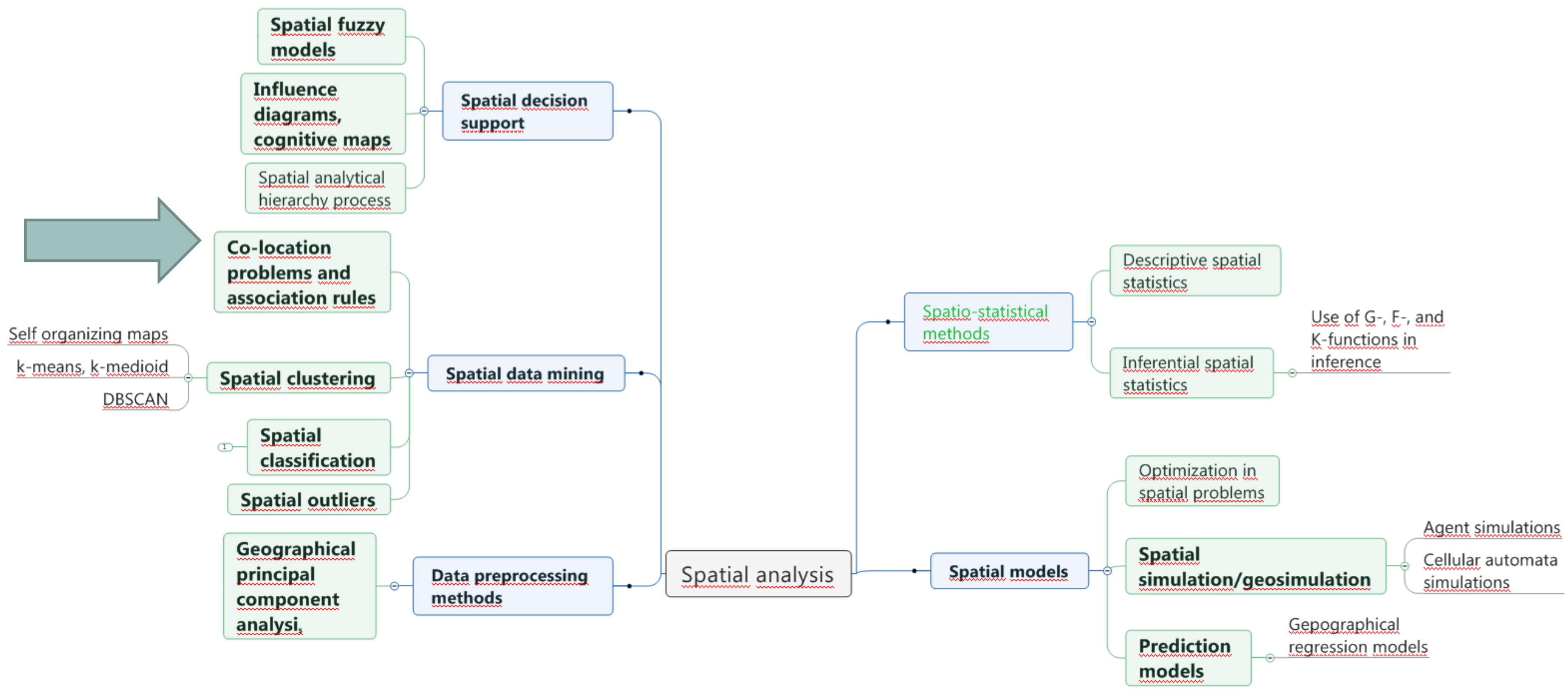
Koperski&Han (1995) have presented a paper on using spatial predicates based on:

- Distance (A and B are close to each other)
- Topology (A and B are adjacent, A and B intersect)
- Direction (compass point) (A is to the North from B)

Association rules are defined and spatial relation are calculated

If data is in a relational database with The Dimensionally Extended nine-*Intersection* Model (DE-9IM) the relationships can be calculated

Not very straightforward, lots of preprocessing and definition of the spatial predicates is required.



Now you have learnt one spatial data mining methods: spatial association rules.

In the following there is an introduction to concepts and methods in spatial data mining.

3. Core concepts in data mining and in spatial DM

Data, information, knowledge

Knowledge discovery = tietämyksen muodostus

Data mining = tiedonlouhinta

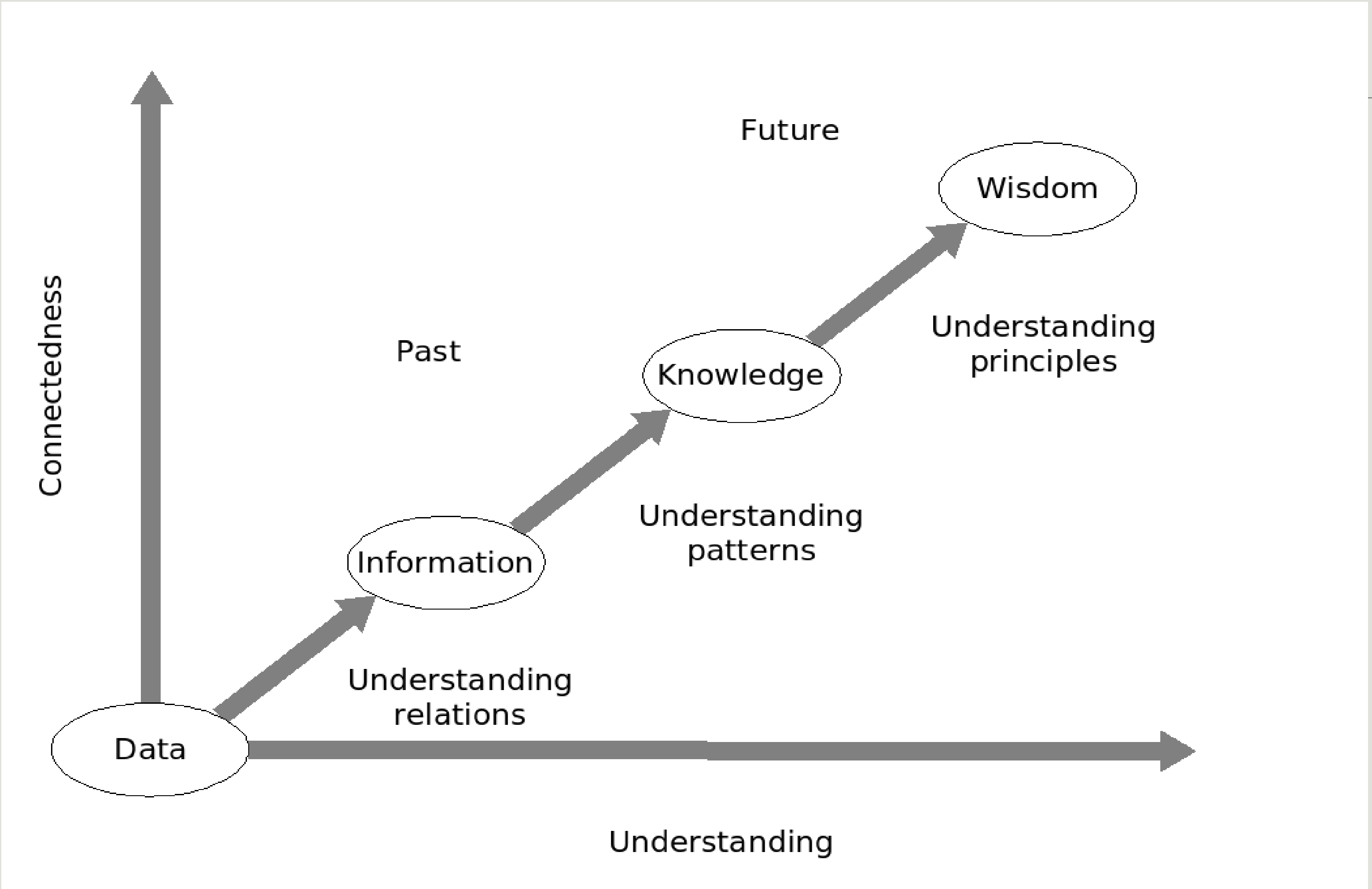
Spatial data mining

Models and patterns = mallit ja kuviot

Spatial relations and patterns

(see also Tiedonlouhinnan sanasto, suomeksi)

Data, Information, Knowledge



Data, information, knowledge
Datasta tietämykseen

data – ”facts”

- not organized, not processed static facts (Awad,2004)

information – ”data in some context”

- data becomes information when it is linked to the context; information has meaning, purpose and relevance (Awad, 2004)

knowledge – ”person’s understanding”

- information becomes knowledge when it is analysed and understood; knowledge is personal, it is based on (personal) perception, skills, education, common sense and experience (Awad, 2004); ”insight” can only be based on knowledge

suomenkielessä ongelma:

- sekä data että informaatio käännetään usein sanaksi ”tieto”
- knowledge = tietämys ; insight = oivallus

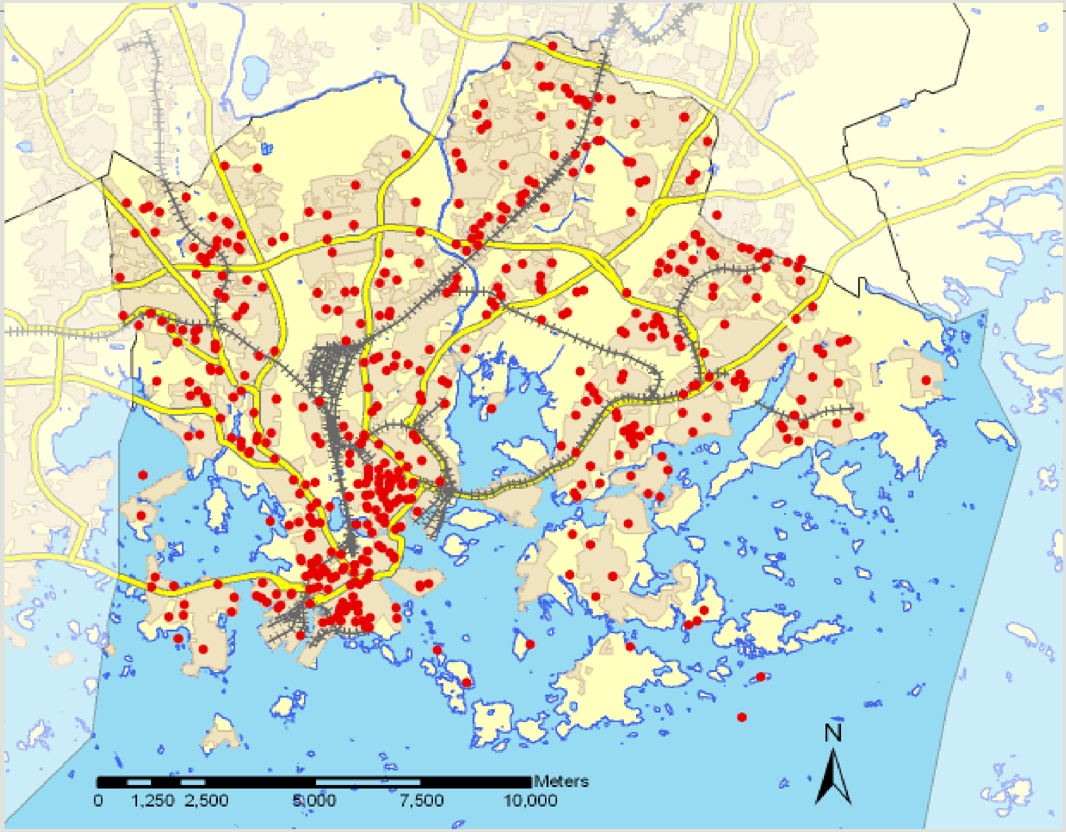
Exercise 2: spatial data, spatial information spatial knowledge

On the following slides you see an example of spatial data analysis case which most of you are familiar with

Try to identify in this analysis process, what are:

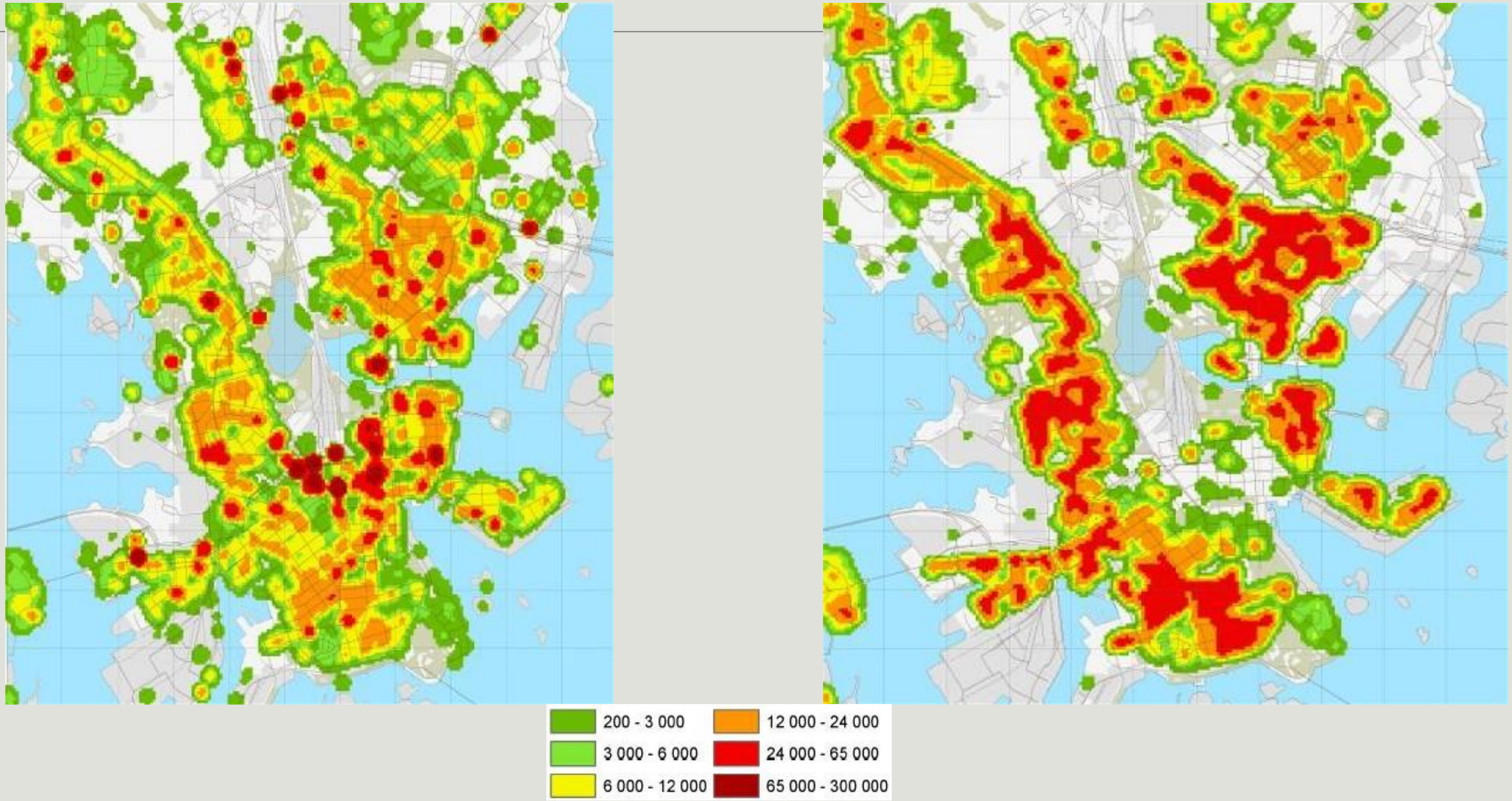
- Data ?
- Information ?
- Knowledge ?
- Wisdom ?

Example: Incidents (domestic fires) in Helsinki City Centre as a Point Pattern; incidents = points

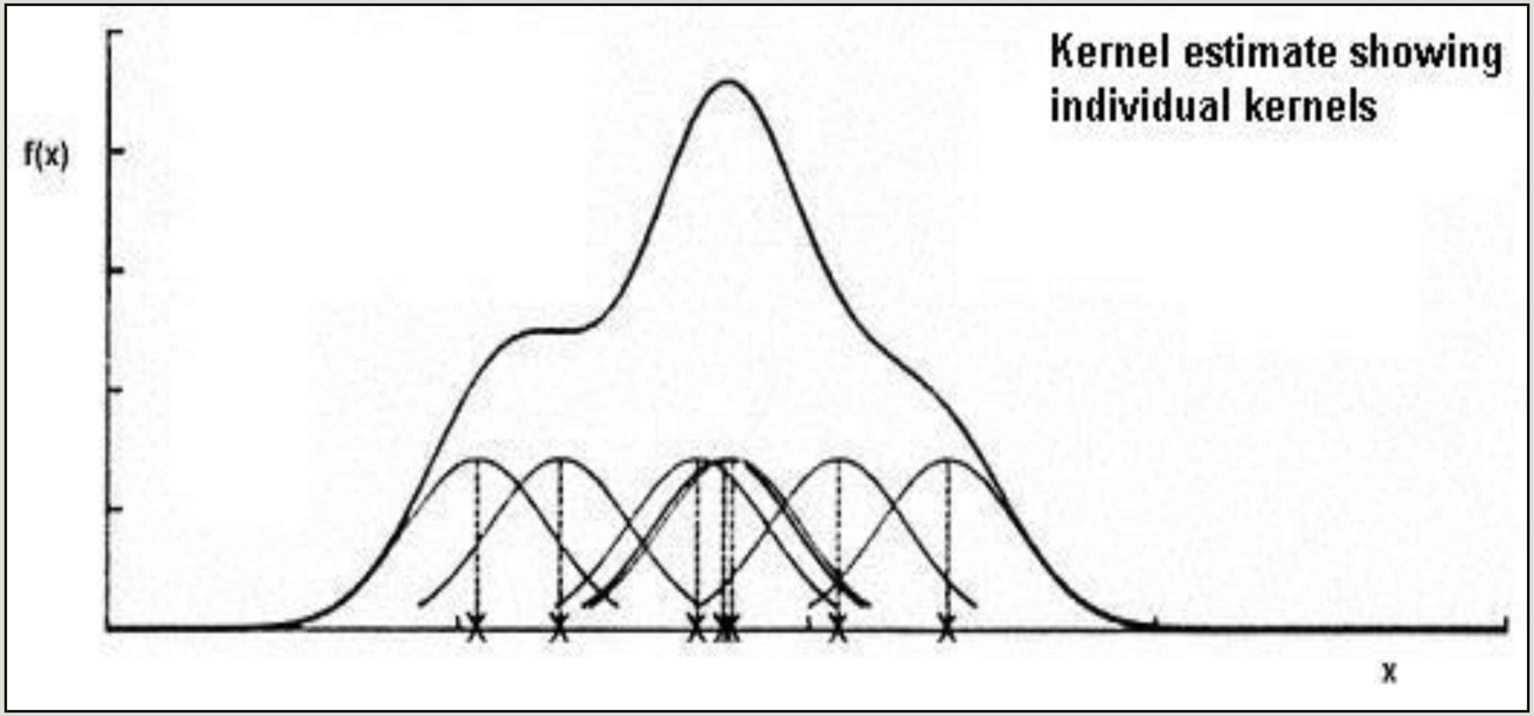


Aalto-yliopisto
Teknillinen korkeakoulu

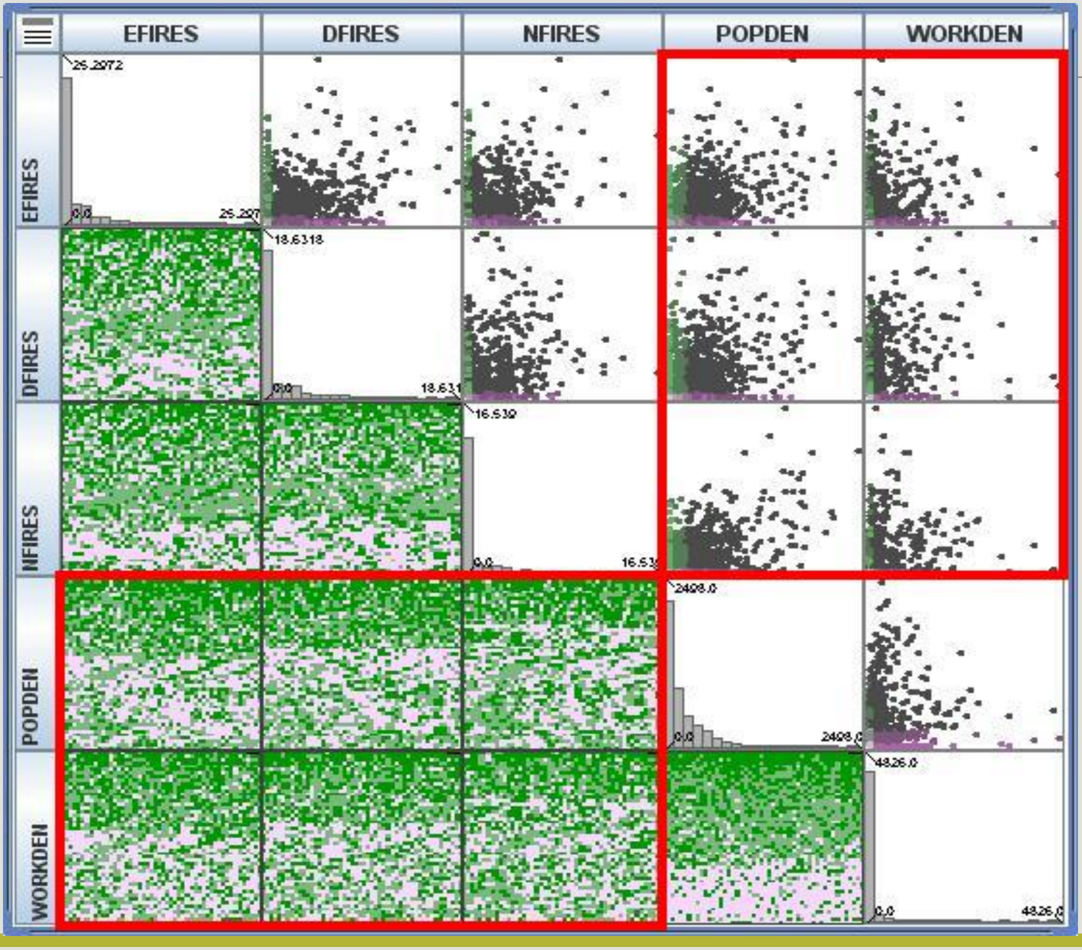
Incidents in Helsinki City Centre, Day and Night – by Kernel density surfaces; difference in distribution



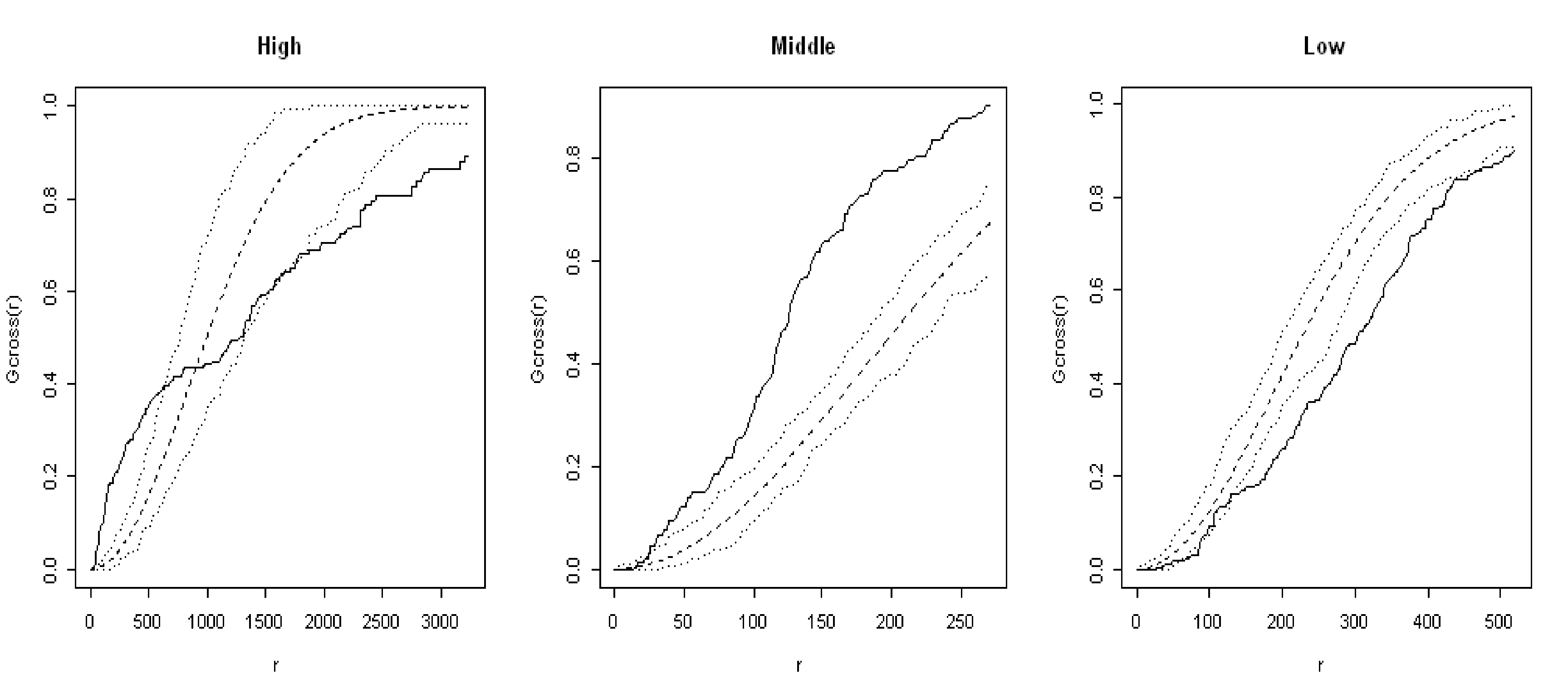
Kernel –density map principle



B-variate matrix: Incidents vs. Population and Workplace density; interpretation : Correlation between night fires and population (Spatenkova, 2009)



Ĝ function for building fires and population density show:
Correlation between two variables

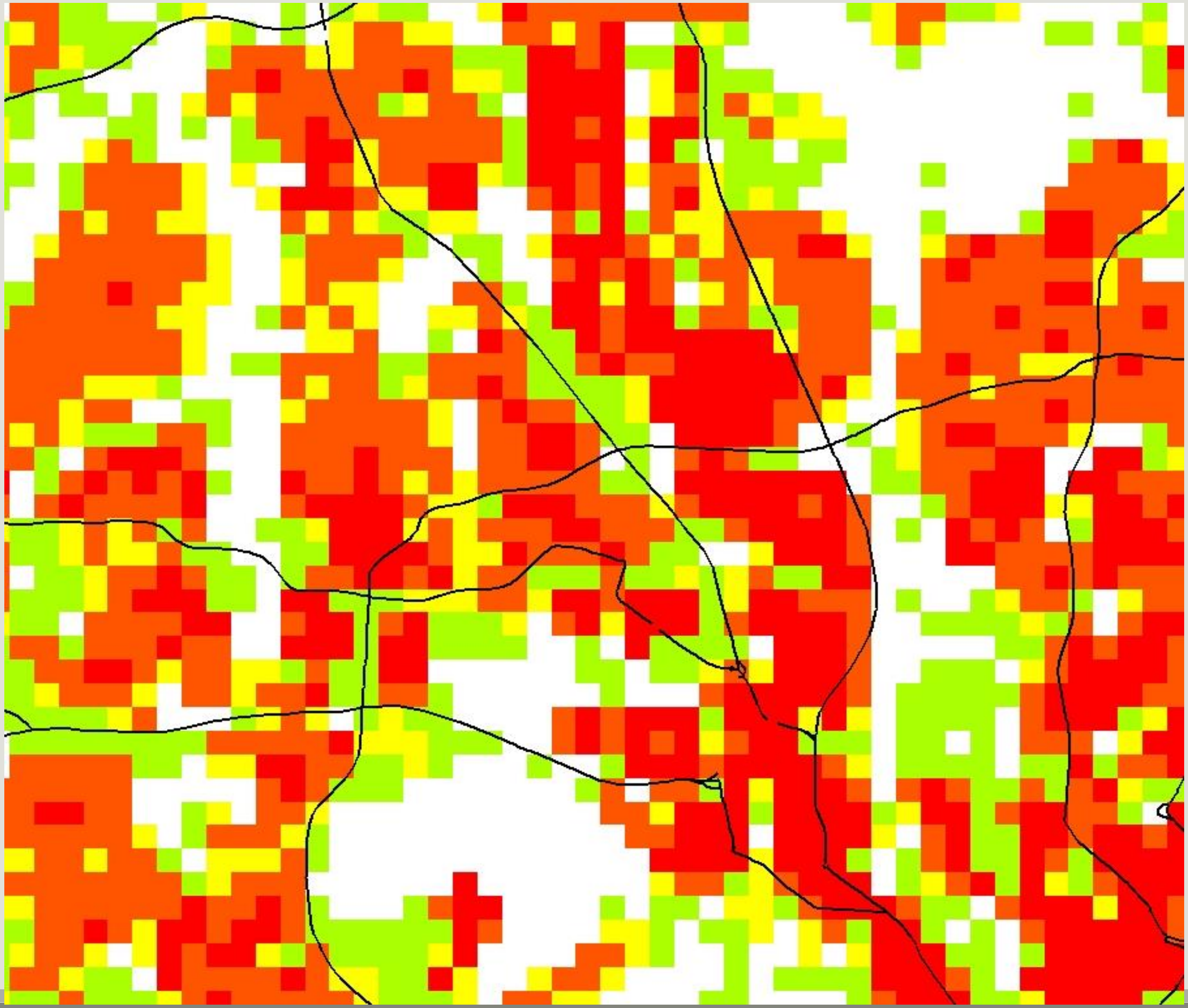


- Ĝ function (solid line)
- theoretical values for random distribution (dashed line)
- simulation envelopes (dotted line)

Spatenkova&Stein,2008)

”Risk level Map” ”Riskitasokartta”

red= high; should be reached in 6 minutes, yellow=medium; 12 minutes, green=low; 20 minutes



Interpretation of this exercise

The goal is to ***analyse*** the phenomenon of building fires

in order to be able ***to identify the relationships*** between fires and some other variables

- correlations, causalities

and ***produce a model*** which can be used for predicting the fires;

- for mitigation and preparedness in the disaster management process/kriisinhallinnan vaiheet: onnettomuuksien/kriisien ehkäisy ja niihin varautuminen
- Risk Level Map of Disasters/ Onnettomuuksien riskitasokartta

the model shows those areas that need to be reached in 6 minutes, 12 minutes, 20 minutes; can be used as a ***decision support*** tool

Relations, patterns and models

Relation (su **suhde)**
- between object types, between phenomena; between attributes
- connection, correlation, dependency

Pattern (su **kuvio, käyttäytymiskuvio)**
- in a specified region of the space, local
- statement about behaviour in restricted region of the space
- example: summary statistics, a simple rule, spatial pattern

Model (su **malli)**
- global
- makes statements about any point in the full measurement space
- example: linear regression model, GWR

Spatial relations, patterns and models

Spatial relations: distance, directional and topological relations

Spatial patterns – often related with the spatial distribution of points, lines and areas together with attribute values:

- **Dense** or **sparse** areas
- **Clusters** based on similarity of data items;
- **Outliers**; data items that appear inconsistent with respect of the remainder of the data set .
- **Classes,** meaningful categories in space;
- Dependency relationships, **associations,** between attributes in the data sets;

Spatial models:

- GWR

Knowledge discovery and data mining

knowledge discovery in databases (KDD) has been defined (Fayyad and Grinstein, 2003) as:

- ***"a process of discovering valid, novel, potentially useful and ultimately understandable patterns from data"***

data mining has been defined by the same authors as:

- ***"the method of extracting patterns from the data"***

Data Mining

definition (Hand, 2001):

- ***"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."***
- *find the difference here to spatial statistics: in data mining there is no hypothesis to test; methods are used without hypothesis, perhaps an intuition about relationship or pattern*

relationships and summaries are

- ***models or patterns*** (linear) equations, rules, clusters, graphs, tree structures, trends, patterns in time series (or in spatial data)

Complete DM process

Includes several subprocesses: (Aggarwal, 2015)

- Data collection (surveys, sensors, other data collection methods into data base)
- Feature extraction and data cleaning, preprocessing
- Algorithm design and tuning, analytical processing
- Analysis of the output

We concentrate on:

- Data mining approcahes/algorithms

Is a nondeterministic and iterative process

The result of the DM process typically is a hypothesis that can then be validated and verified by statistical methods

Spatial Data Mining

Spatial data mining is a knowledge discovery process of extracting implicit interesting knowledge, spatial relations, or other patterns not explicitly stored in databases. [Koperski et al. 1996]

”Spatial data mining is the process of discovering interesting and previously unknown but potentially useful patterns from spatial databases.” (Shekhar, 2011)

The goal of SDM

…is ”to discover interesting and potentially useful patterns of information embedded in large databases”

…”the goal of SDM is to automate the discoveries of such correlations which can be then examined by specialists for further validation and verification” (Shekhar&Chawla)

The core question actually is:

- what is interesting, what we are looking for ?

In SDM basically we do not ”know” in the beginning exactly what we are looking for, but at some step of the process we have to agree about the problem what we are solving, see the Fig 7.1 , p. 184 in Shekhar&Chawla

Computational SDM methods

vs. spatio-statistical analysis ? – what is the difference?

- statistical ideas and methods are fundamental to data mining

differences

- the ***size of data sets***; in statistics sampling is used
- mining quite often uses ***data collected for other purposes***; statistical methods typically use specially collected data sets
- mining algorithms can use also ***non-complete data*** sets (missing data)

4. How the generic methods of data mining can be applied to spatial data ?

- spatial and spatio-temporal data are more complex than non-spatial data, different data types

- also the mining methods have more challenges

- spatial data have special features:
  - many dimensions: 2d,3d,4d + attributes
  - object and field data models; vector and raster implementations
  - all data is not just attributes in multidimensional data bases
  - graph structured data, trajectories

coordinates and metric relationships: distance, direction

topological relationships: adjacency, connectivity

tendency to spatial autocorrelation

spatial heterogeneity

From data mining to spatial data mining

- In this course you learn some data mining methods that has been developed into spatial ones by adding some method for **management of the spatial heterogeneity and dependence**

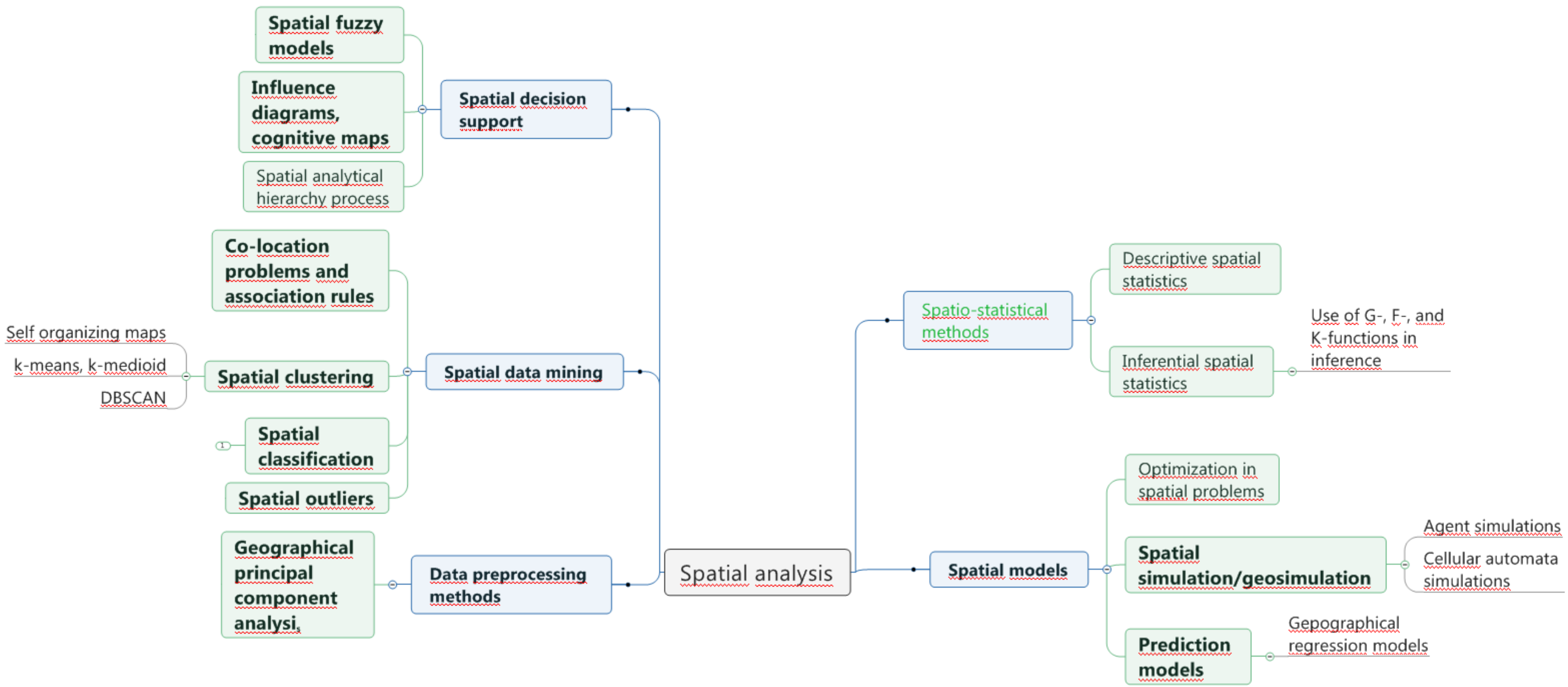
-Methods that can make data mining methods into spatial

Association rules – **spatial predicates and indexing/spatial structuring**

Clustering – **location as one of the attributes**

Classification – **SAR and CAR models; use of gamma-index and focal method; W-matrix**

Geographically weighted regression, (interpolation, density estimation) **– Kernel weighting**


Spatial analysis
Spatial decision support
Spatial fuzzy models
Influence diagrams, cognitive maps
Spatial analytical hierarchy process
Spatial data mining
Co-location problems and association rules
Spatial clustering
Self organizing maps
k-means, k-medioid
DBSCAN
Spatial classification
Spatial outliers
Data preprocessing methods
Geographical principal component analysi,
Spatio-statistical methods
Descriptive spatial statistics
Inferential spatial statistics
Use of G-, F-, and K-functions in inference
Spatial models
Optimization in spatial problems
Spatial simulation/geosimulation
Agent simulations
Cellular automata simulations
Prediction models
Gepographical regression models

4.1 Spatial predicates

Spatial data is special in the sense that data has locations and while having locations also spatial relations exist.

Spatial relations can be distance, directional or topologic

> Example: *a country that is adjacent to the Mediterranean Sea is a wine-exporter*

Spatial relationships can be described by spatial predicates

Spatial predicates have been standardized (for data base management use) by using the so-called 9-intersection model

By spatial predicates various relationships can be expressed for spatial data mining purposes

4.2 Spatial index or spatial buffer

A simple way of managing spatial distance relationships is to define neighbourhood concept and measure then co-location

The methods to identify co-location:

- use some kind of spatial index; like grid
- or create some neighbourhood areas; circles of even Voronoi polygons

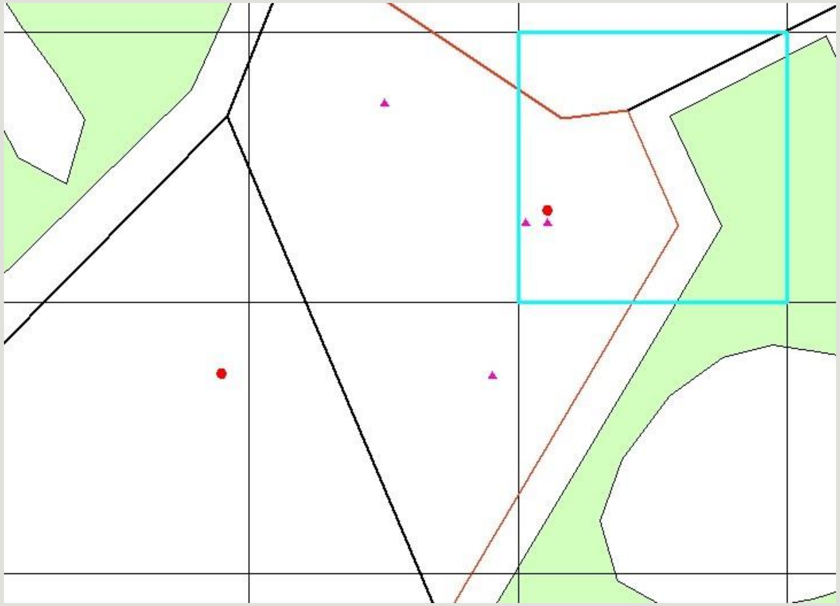
The search of co-location is then based on this indexing

This method is very simple and has lots of limitations

Spatial indexing for co-location

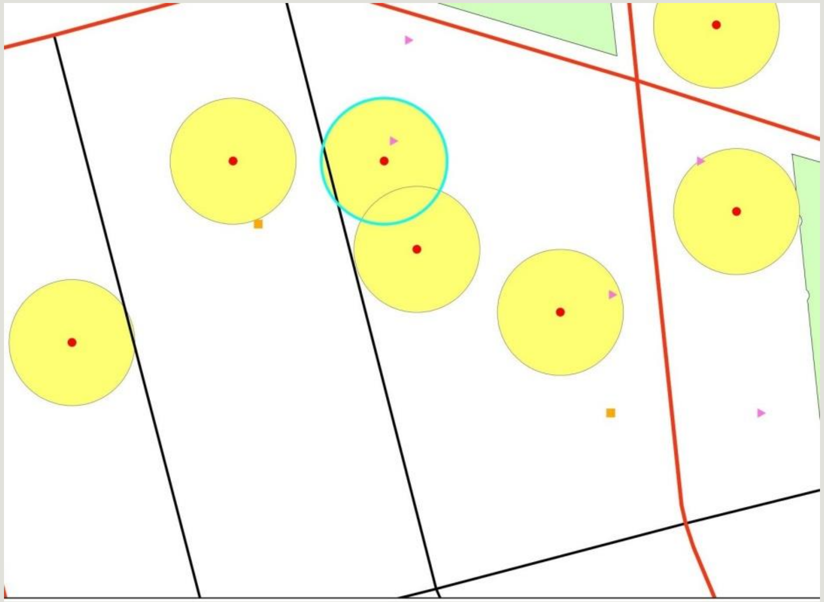
Grid approach

square regular grid over the whole study area (50 x 50 m)



Buffer approach

- circular buffer around incidents (r = 50 m)



Spatial predicates and spatial co-location by index/buffer in spatial association rules

A simple way of analysing interaction between variables is to analyse **relationships within attributes in a relation**

The method is known as discovery of association rules

Association rules is maybe the simplest data mining technique

Spatial predicates and concept of co-location is used in developing association rules into spatial method

Apriori –algorithm is the most well-known algorithm for assciation rule discovery

Example: *high co-existence of a bar or restaurant and an incident in a geographic location* is a typical association rule

4.3 Dividing the study area into subareas

If the phenomenon in question is known so far that the study area can be divided into meaningful subareas, this can be one method to manage spatial heterogeneity

Subareas can be selected according to the density of the objects (in case of population of municipalities) or according to the directional spatial behaviour (in case of anisotropy in the data set)

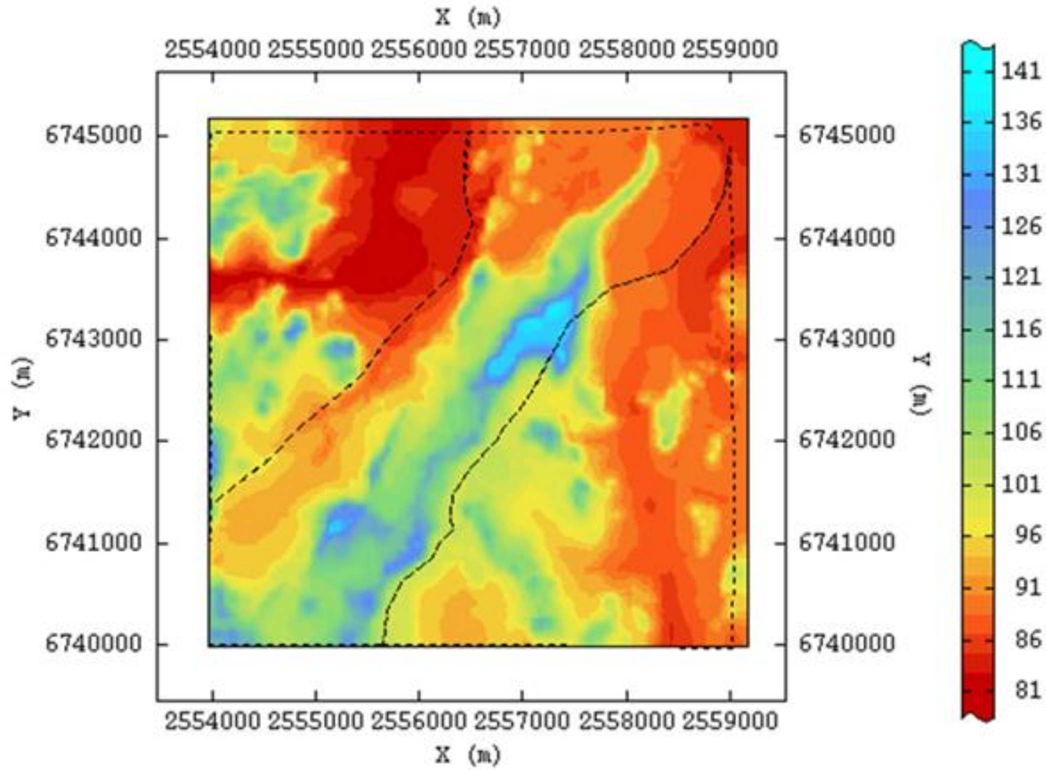
Autocorrelation models in Kriging

In Kriging semivariogram method is used for revealing the possible spatial autocorrelation

Semivariogram models gives the required parameters for Kriging interpolation

Kriging has some challenges for modeling directional differences in autocorrelation; sometimes the study area must be divided into subareas in order to get realistic results

Kriging in 3 areas (from Rangsima Sunila's slides)



4.4 Methods for add spatial homogeneity in the method

Post-processing after **non-spatial data mining**

Using **coordinates as attributes or some other coordinate based measure** in **the data mining computation**

- Is not as straightforward as it looks

How to use clustering for spatial data ?

Clustering is one of the most popular data mining methods

The idea of clustering is to analyse the similarity of objects/pixels by calculating the distance in multivariate space

Various algorithms exist for finding the most similar groups of objects

- For example k-means
- In spatial clustering the problem is that also **the location means**

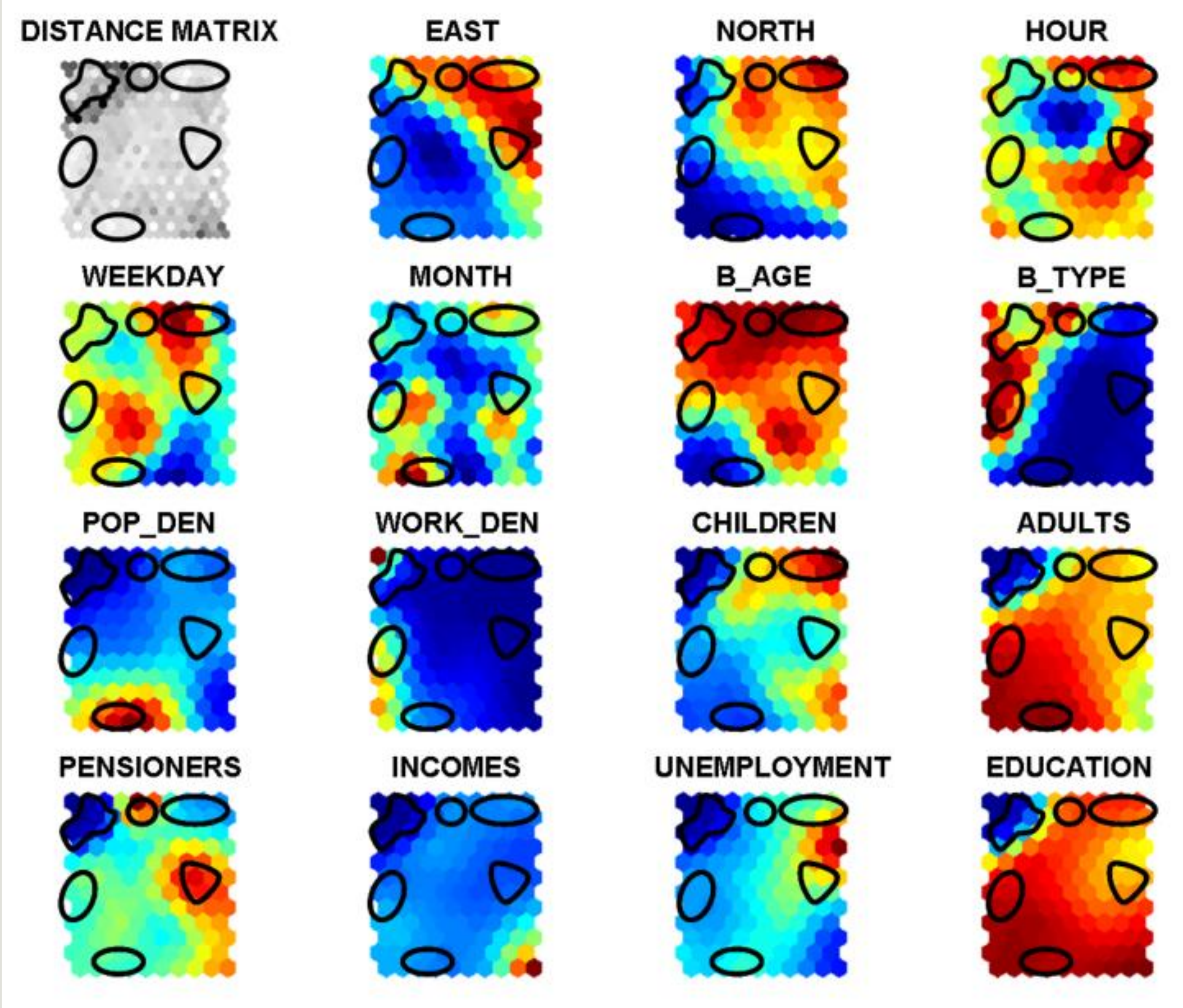
Challenge of spatial clustering:How to produce groups/clusters that include similar objects which also are close in geography ?

Some examples

Algorithms that are based on geographical distance

- Incorporating **spatial contiguity constraints in hierarchical clustering** method; the result is homogeneous units; source
- **Example: Regionalization of forest pattern metrics for the continental United States using contiguity constrained clustering and partitioning**
- John A.KupferPengGaoDianshengGuo

- In using SOM (self organizing map) coordinates or distances can also be incorporated as attributes (Spatenkova, O., 2009)



Coordinates as attributes; SOM by Špatenková (2009)

Attribute can also be distance to river, closest building etc.

Attributes can also be calculated using location, for example distance to city centre.

4.5 SAR and CAR in spatial classification

Spatial autocorrelation can be included to spatial models like regression or simulation models by adding there a component of spatial autoregressive model

- In dependent variables of regression model, SAR model
- By using Bayes conditional methods, CAR model

These methods are used in data mining methods in order to get better results for prediction and identifying dependencies between variables

To avoid ”salt and pepper” effect

SAR models – spatial lag

http://www.statsref.com/HTML/index.html?sar_models.html

A pure SAR model consists of a lagged version of regression model

Idea: The dependent variable is dependent on the values of neighbouring locations

$$X = \rho W X + \varepsilon$$

$$X = (I - \rho W)^{-1} \varepsilon$$

W-matrix that contains adjacency information

Rho that stands for the strength of autocorrelation

CAR models

http://www.statsref.com/HTML/index.html?car_models.html

produces similar results than SAR

Idea: the probability of values estimated at any given location are conditional on the level of neighboring values

the form of CAR:

$$E\left(y_i \mid all\ y_{j\neq i}\right) = \mu_i + \rho \sum_{j\neq i} w_{ij}\left(y_j - \mu_j\right)$$

where μ_i is the expected value at i, and ρ is a spatial autocorrelation parameter, w is the adjacency matrix

the formula can be used in form of spatial decay, when the strength of spatial autocorrelation must be analysed for example by using semivariogram

Spatial classification

Many problems can be categorized as classification problems, for example

- Location prediction or thematic classification

- in many cases **spatial autocorrelation** exist in the data set and neighbouring pixels belong more likely to same class than pixels with longer distance between

Example of location prediction:

- to predict whether an event occurs in a geographical location, or not, based on the analysis of other socio-economical data; **regression models** can be used in solving the relationships between independent an dependent variable

Example of thematic classification:

- to categorize all geographical locations into as good classes as possible;

4.6 Using neighbourhood similarity analysis – Focal method for spatial autocorrelation

Adjacency matrix W carries information about entity values, for example classes in grid structure

Concepts used:

- Adjacency matrix W; describes the classification structure – equal class in neighbourhood marked with 1
- Focal autocorrelation statistic, Gamma index is used; Gamma index is calculated based on W and when the value of Gamma is negative the focal situation is most probably salt-and-pepper; this measure is usedin sio-called focal text and example of a method using this is **spatial decision tree**

4.7 Modelling spatial heterogeneity by geographically weighted methods

Spatial heterogeneity exists when the structure of the process being modeled varies across the study area

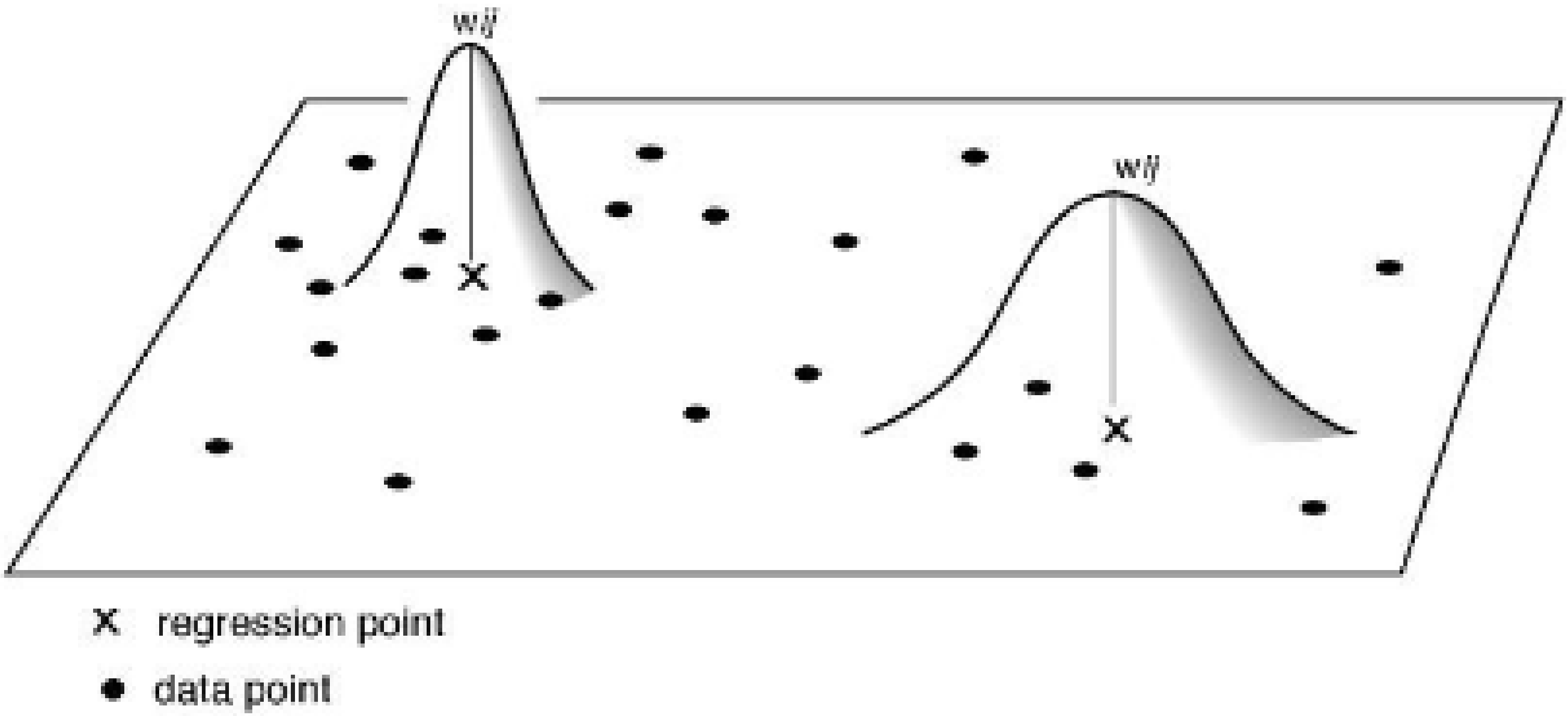
For example in linear regression models the relationship between variables is constant over the space and residuals (error term) are normally distributed (show no autocorrelation)

Global models do not always fit and Geographically weighted models give better results

Spatial heterogeneity is modeled by using Kernel function for weighting the variables; kernel bandwidth changes over the space

The same method can be used in many other mathematical methods

Spatially Adaptive Weighting Scheme



Stewart Fotheringham

Reading material for this lecture

Zhe Jiang, Shashi Shekhar, Spatial Big Data Science – Classification techniques for Earth observation technology, 2017. Chapters 1 and 2

- overview on spatial data mining and spatial autocorrelation and heterogeneity

Shekhar,S., Evans,M., Kang,J.,Mohan,P., Identifying patterns in spatial information: a survey of methods. 2011

Shekhar,S., Chawla,S., Spatial Database Book, Prentice Hall, 2003. Chapter 7 you can download from www.spatial.cs.umn.edu/Book/

- These two materials give an overview on spatial data mining and also introduction to the tpics of this lecture (association rules)

Literature

Shekhar,S., Evans,M., Kang,J.,Mohan,P., Identifying patterns in spatial information: a survey of methods. 2011.

Shekhar,S., Chawla,S., Spatial Database Book, Prentice Hall, 2003. Chapter 7 you can download from www.spatial.cs.umn.edu/Book/

Geographic Data Mining and Knowlededge Discovery, edited by Miller,H.J. and Han, J., 2001.

Aggarwal,C., Data mining, 2015.

Zhe Jiang, Shashi Shekhar, Spatial Big Data Science – Classification techniques for Earth observation technology, 2017.

Hand, D.,Mannila,H., Smyth,P., Principles of Data Mining, 2001.

Miller,H., Geographic data Mining and Knowledge Discovery, in Hadbook of GIScience by Fotherinham et al.

Spatenkova,O., Discovering spatio-temporal relationships: A Case study of risk modelling of domestic fires, doctoral dissertation, HUT, 2009.

Karasova,V., Spatial data mining as a tool for improving geographicxal models, M.Sc thesis, HUT, Dept. of Surveying, 2005.

Demsar,U., Exploring geographical metadata by automatic and visual data mining, Lic.thesis, KHT, Stockholm, 2004.

Kovalerchuk,B., Schwing,J., Visual and spatial analysis; Advances in data mining, reasoning and problem solving, 2004