HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Spatial Data Mining

Antti Leino ‹antti.leino@cs.helsinki.fi›

Department of Computer Science

Overview

- Spatial Data Mining
  - Exploratory methods for analysing data
  - Spatial component
  - Emphasis on point data
- Main topics
  - Co-location rules
  - Spatial clustering
  - Spatial modelling

Administrivia

- Lectures / meetings
  - 12th March – 26th April 2007
  - Mon, Thu 10–12 am, C222
  - Introductory lecture for each main topic
  - Other times two articles / meeting
    - Presentation by a student, c. 20 min
    - Discussion
- Exam Thu, 3rd May, 4–7 pm
- Project work by Wed, 16th May
  - Exercise in spatial data mining
  - Essay on a related topic
  - Course diary
- http://www.cs.helsinki.fi/u/leino/opetus/spatial-k07/

Schedule

12.3. Introduction

15.3. Co-location patterns

19.3. Huang & al., 'Discovering Colocation Patterns from Spatial Data Sets: A General Approach'
- Salmenkivi, 'Efficient Mining of Correlation Patterns in Spatial Point Data'

22.3. Yoo & al., 'A Joinless Approach for Mining Spatial Colocation Patterns'
- Huang & al., 'Can We Apply Projection Based Frequent Pattern Mining Paradigm to Spatial Colocation Mining?'

Schedule

26.3. Xiong & al., 'A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects'
- Yoo & al., 'Discovery of Co-evolving Spatial Event Sets'

29.3. Spatial clustering

2.4. Tung & al., 'Spatial Clustering in the Presence of Obstacles'
- Wang & Hamilton, 'DBRS: A Density-Based Spatial Clustering Method with Random Sampling'

– – Easter break

Schedule

16.4. Spatial modelling

19.4. Kavouras, 'Understanding and Modelling Spatial Change'
- Kazar & al., 'Comparing Exact and Approximate Spatial Auto-Regression Model Solutions for Spatial Data Analysis'

23.4. Shekhar & al., 'A Unified Approach to Detecting Spatial Outliers'
- Hyvönen & al., 'Multivariate Analysis of Finnish Dialect Data – an overview of lexical variation'

26.4. Summary

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

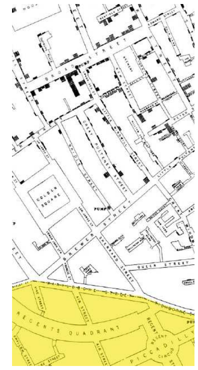
Spatial Data Mining Introduction

Antti Leino ‹antti.leino@cs.helsinki.fi›

Department of Computer Science

Introductory example: Lontoo 1854

- 1854 cholera epidemic
- Hero of the story: John Snow, MD
- Method: plot on a map
  - Cholera deaths
  - Public water pumps
- Discovery:
  - Deaths cluster around one pump
- The epidemic was shut down by removing the handle from the pump



Rest of the story

- Snow 1849: theory that cholera is transmitted by polluted water
  - The data mining experiment a part of testing the theory
- London had two water companies
  - One took its water from the Thames upriver of town, the other downriver
  - The polluted pump belonged to the latter
- Follow-up studies to confirm that
  - The cholera victims had used the polluted pump
  - Those who did not use the pump did not fall ill
  - In other words, results were verified by other means

Nevertheless . . .

- Rather low impact
  - The episode involved only one district in London
  - The polluted pump was reopened some weeks later
  - Snow's theory was finally accepted a couple of decades later
  - Snow became famous in 1936
- Hindsight is easy
- Classic examples often have mythical elements

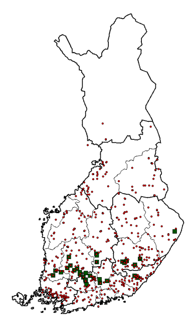
Data Mining

- Extract new, interesting information from massive amounts of data
- New
  - Surprising
  - Not too strict prior expectations
- Interesting
  - Relevant, useful
  - Often requires some knowledge of the application
- Spatial data mining: add a spatial component

Different kinds of data

Point patterns

- Shape is not relevant
- Each phenomenon represented by a separate point pattern
- Example: Viking-age forts
  - Red dots: place names starting with *Linna*- 'castle'
  - Green squares: Viking-age hill forts

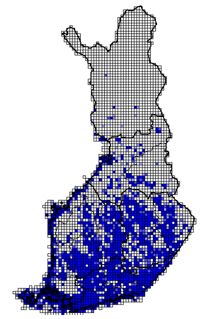


Different kinds of data
Spatially continuous data

- Describes a spatially continuous phenomenon
- Not possible to measure across the space
  - Measurements at distinct points
- Measurement points not interesting as point patterns
- Goal: model the phenomenon in order to predict the values between the measurement points

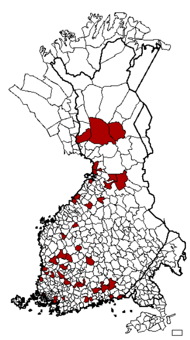
Different kinds of data
Area data

- Spatial variation presented as regions
- Example 1: spatially continuous phenomenon
  - Breeding certainty of the great crested grebe
  - Finland divided into 10×10 km squares



Different kinds of data
Area data

- Example 2: Distinct area
  - Spatial distribution of a dialect word *Aaholli*
- Somewhat like a point pattern, but now the shape is meaningful

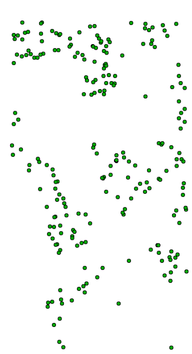


Co-location rules

- Typically for point patterns
  - Correlation between different point patterns
  - 'Members of these point patterns often occur close to each other'
- Similar correlations can be established for area data
- Spatial association rules
  - 'If phenomena $A_1,\ldots,A_n$ are found near each other, phenomenon B is also likely to be found'
- Cf. frequent sets and association rules in transaction data

Spatial clustering

- Goal: find clusters in a point pattern
  - Areas with high point density
  - Separated by areas with low density
- Example: farms
  - Green dots: farm locations
  - Large-scale clustering: areas divided by lakes
  - Smaller-scale clustering: villages



Spatial modelling

- Generally: find a model that
  - describes the phenomenon
    - Underlying factors or variables
  - can be used for predictions
    - Areas that have not been surveyed
    - Effect of changes
- Two phases
  - Select a suitable model
  - Find the parameters for the model
- Example: dialect words
  - Principal component analysis