HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Computational analysis of spatial co-location rules

Antti Leino ⟨antti.leino@cs.helsinki.fi⟩

GI Norden, 4th October 2006

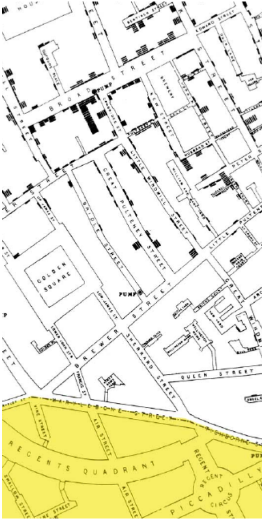
Department of Computer Science
Research Institute for the Languages of Finland

Data Mining

- Sub-field in computer science
- Goal: find interesting new information in a large collection of raw data
  - Interesting
    - Relevant
    - Useful
    - Requires knowledge of the field
  - New
    - Surprising; not obvious
    - Few a priori notions

Background
London 1854

- Cholera epidemic
- Hero: Dr. John Snow
- Method: plot on map
  - cholera deaths
  - public water pumps
- Discovery: deaths cluster around one pump
- Solution: remove the handle from this pump



London 1854
Continued

- Snow 1849: theory that cholera is transmitted via polluted water
  - the spatial analysis a part of testing this

- London had two water companies
  - One took its water from the Thames above the city, the other below
  - The polluted pump belonged to the latter company

- Subsequent study to make sure that
  - Cholera victims used the polluted pump
  - People who didn’t use the pump did not fall ill
  - That is, the results were confirmed

London 1854

But

- Not widely accepted at the time
    - Only one region in London
    - The polluted pump was reopened after a few weeks
    - Snow’s theory eventually accepted a couple of decades later
    - Snow’s fame stems from 1936
- Classic examples often have mythical elements

Co-location patterns in names

From statistics to onomastics

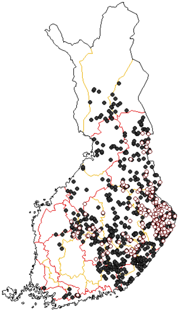
- Starting point: Place Name Register
  - National Land Survey
  - Part of the Geographic Names Register
  - All names on the 1:20 000 basic map
  - Each named place presented as a point

- What can one do with this?

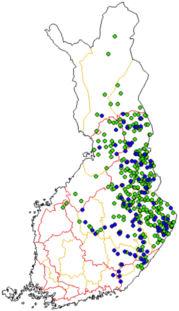
Co-location patterns in names

Maps

- Names in each pair have roughly similar distributions
- Not easy to see whether they attract each other



Mustalampi
'Black Pond'
Valkealampi
'White Pond'



Kuikkalampi
'Diver Pond'
Ruunalampi
'Gelding Pond'

Co-location patterns in names

Spatial statistics

- A place name has a distribution
  - Can be considered a (marked) point pattern

- The K function

- $K(r)\lambda = E($number of points within radius r of a random point$)$
  - λ overall the intensity of points

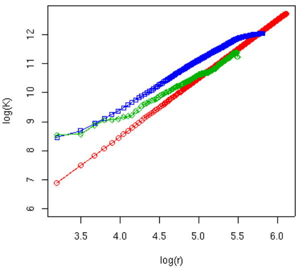
- $K(r) \approx$ the area around a point which one would need to expect the actually observed number of points

Co-location patterns in names

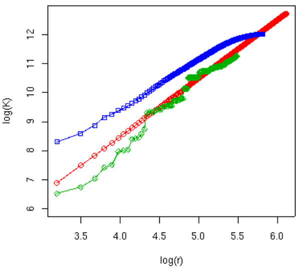
Spatial statistics

- The *K* functions look similar
- Substitute the uniform λ with a dynamic $\lambda(s)$
- Now the pairs are different!

Mustalampi - Valkealampi



Kuikkalampi - Ruunalampi



Co-location patterns in names

Data mining

- Find pairs whose cross-*K* function indicates attraction
- Join these into larger groups
- Use these as the basis for further analysis

- Other ways to mine co-location patterns
  - Many are more effective than this
  - Most have potentially problematical assumptions, such as a uniform intensity
  - Choose the right tool

Co-location patterns in names

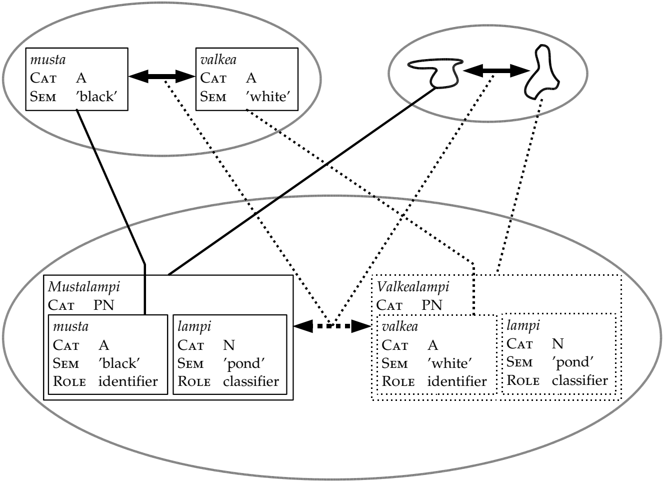
Onomastics

- These groups of names have interesting implications
  - Contrastive names quite common
  - Naming process often based on such contrast
  - Meaning of name elements important
  - Interplay between the meaning of the elements and the referents of the names
  - . . .

- In other words, exploratory data analysis only first step
- Starting point for further linguistic analysis

Co-location patterns in names

Onomastics



Co-location patterns in names

Onomastics

- Previous slide showed
  - Name structure using the formalism of Construction Grammar
  - Pattern-based naming process in terms of conceptual blending

- Getting here required
  - Place Name Register
  - Spatial data mining
  - Onomastic analysis

Summary

- Knowledge discovery is a long process
- Elements from several fields
  - Statistics
  - Data mining
  - Application fields
- There is a lot of spatial data
- Mining it is useful

Thank you