Introduction to Spatial Analysis and Spatial Data Mining

Kirsi Virrantaus
Department of Built Environment
School of Engineering
Aalto University

Introduction to Data Science Module
Courses on Spatial Analytics and Spatial Data Science offered in the Module (2016-2017)

• GIS-E1060 Spatial Analytics, 5 cr, II
  – introduction to the statistical and visual methods for analysis of spatial data
  – lectures and computer class exercises

• GIS-E4020 Advanced Spatial Analytics, 5 cr, V
  – introduction to the spatially extended data mining methods
  – spatial classification, spatial clustering, spatial association rules, mining graph data from physical networks, trajectory/moving objects data mining
  – methods, applications and introduction to algorithms
  – lectures and computer class exercises
Contents and learning goals of this lecture

1. What is so special in spatial phenomena and data?
   – you will understand the **special requirements of spatial data** from analysis point of view

2. Spatial autocorrelation
   – you understand the concept of **spatial autocorrelation, hot spot and spatial cluster**
   – you can describe the most **popular methods** to identify the type of spatial distribution of the data set

3. Spatial data mining
   – you can describe the **methods** required to extend non-spatial data mining methods to be usable for spatial data

4. Integrated methods
   – you understand the **power of visualization** and use of **human knowledge** in decision support
Visualization of slope and aspect of digital elevation model. DEM is modeled in grid structure, slope and aspect calculated from the data.
Soil map is a typical continuous phenomenon modeled in grid structure and classified according to predefined class definitions (Sunila, 2009).
Incidents stored in the register of Fire and Rescue Helsinki, as a dot map.

(Spatenkova, 2009)
A subset of the incidents presented as Kernel density surface. Visualization emphasizes the clustering effect.

(Krisp, 2006)
Roads and streets of Helsinki from The Digiroad Data Base

ZhangZhe, 2010
1. Spatial is Special

"Spatial data is special because:

- it is **multidimensional** – 2, 3, 4 dimensions (space+time) + n attributes
- it is **voluminous** – easily reach a terabyte in size
- it must be **projected on plane** – in order to be presented
- it requires special (extended) **methods for analysis**
- integration of several data sets can be time-consuming
- updating processes are complex and expensive
- display of maps requires retrieval of large amounts of data” (Longley et al.)

- Spatial data requires special **methods for modeling, storing, processing and analyzing efficiently and correctly**
Example: Fire and Rescue Data

• For example phenomenon ”Incidents in Helsinki” can be described as a dot map in which each dot represents an incident (accident)
• By viewing the map it can be seen that this data is not randomly distributed and the assumption of complete spatial randomness can not be made for data analysis purposes
• Independent random process (IRP), complete spatial randomness (CSR), means
  – Condition of equal probability
    • Any point has equal probability of being in any position
  – Condition of independence
    • Positioning of any point is independent of the positioning of any other point
• The spatial behaviour of this phenomenon can not be modeled or analysed by straightforwardly using regular distributions and methods
2. Spatial autocorrelation

- Spatial data sets describe **spatially (and temporally) distributed phenomena**
- The distribution of spatial phenomena are affected by various **physical obstacles** (roads, rivers, lakes, sea)
- Spatial phenomena have also special intrinsic characteristics like: **spatial autocorrelation**
- Tobler’s 1st Law of Geography says:
  - ”nearby things seem to be more similar than things further away” = spatial autocorrelation
- Spatial phenomena can not be assumed to behave randomly, but **assumed to autocorrelate**
- And also **to be dependent on other** phenomena in the spatio-temporal space
Incidents stored in the register of Fire and Rescue Helsinki, as a dot map.

Incidents are not located anywhere - not at sea - not in places without inhabitants

(Spatenkova, 2009)
Concepts used

• Hot spots, cold spots
  – In hot spots there is more events that in average (high values)
  – In cold spots there is less events than in average (low values)

• Spatial clusters
  – Spatial clusters are groups of similar type of events which also are located close to each others
Cells of specific soil type are clustering into bigger soil polygons.
3. Density based methods to identify hot spots and clusters

- Density based methods can be used for identifying hot spots and clusters

- Popular methods:
  - Kernel density method: visual methods to identify clusters
  - Quadrat count method: simple statistics to identify clusters
Example: Incidents (domestic fires) in Helsinki City Centre are not distributed regularly or randomly.
Incidents in Helsinki City Centre, Day and Night – different clusters depending on the time of the day

Spatenkova, 2009
Kernel-method – visualization

(Krisp, J.)

- Kernel method
- from discrete observations to surfaces

(Krisp, 2006)
Quadrat method – simple statistics

- so-called **quadrat methods**
  - the region is divided into subareas
  - amount of events in each quadrat are recorded
- the quadrats can fill the study region with no overlaps
- the quadrats can be randomly placed
- we can compute
  - quadrat counts – number of events in each quadrat
  - frequency distribution
- binomial distribution – or the more practical Poisson distribution is the null hypothesis of the point pattern
  - if variance/mean(VMR) = 1, distribution is Poisson
  - if the ratio > 1, the point pattern is more clustered
  - if the ratio < 1, the point pattern is more evenly distributed
Distance based methods for identifying clusters and spatial correlation

- Distance from each point to the nearest neighbour
- Distances between every possible point pair

- Popular methods
  - **G-, F-, and K-functions**: distances to nearest neighbours calculated and presented as a cumulative frequency graph; clustering or not
  - **Variogram cloud**: distances between all point pairs are calculated and presented as a point cloud; spatial autocorrelation or not

- Human interpretation based on visual analysis
\( \hat{G} \) function showing correlated building fires and population density

- \( \hat{G} \) function (solid line) (empirical data)
- Theoretical values for random distribution (dashed line) (model)
  - Simulation envelopes (dotted line)
Variogram cloud

• reflects the relationships between attribute values and spatial location of the entities in the data set

• examples in the following figures (2.6 and 2.7; O´Sullivan&Unwin)
  – terrain as a contour map (2.6)
  – same data organized as a variogram cloud: for all possible point pairs the square root of the difference in their heights is plotted against the distance they are apart
  – interpretation of the variogram cloud tells about the autocorrelation in the data set
Spot heights and their contour pattern. Note that this contour pattern is an approximation only and was done by hand.

(O’Sullivan & Unwin, 2003)
Figure 2.7  Variogram cloud for the spot height data in Figure 2.6.
Figure 2.8  Variogram clouds for N–S oriented pairs in Figure 2.6 (open circles), and for E–W oriented pairs (filled circles).
• analysis of N-S and W-E oriented pairs
  – in the Fig 2.8 we can see that the differences between N-S oriented pairs are greater than between E-W oriented pairs; matches with the contour map
  – this phenomenon is called as anisotropy = there is a directional effect in the spatial variation of the data
4. Spatial data Mining

- Use of non-spatial data mining methods
  - classification, clustering, association rules
- Adding required extensions to manage
  - spatial autocorrelation
  - other requirements of spatial data and phenomena
- Use of special spatial data mining methods
  - Trajectory data mining, graph/network data mining
Managing spatial characteristics

• Spatial regression model
  – the simplest way to modify the regression equation is to use the contiguity matrix/adjacency matrix $W$
  
  the spatial autoregressive regression equation is then
  
  $$Y = \rho W Y + X\beta + \varepsilon$$

  – in which $\rho$ is the parameter that shows the correlation strength
  – when $\rho$ is 0, the equation collapses to the classical regression model

• Spatial association rules
  – Rules are defined in terms of spatial predicates, not items
  – Example: a country that is adjacent to the Mediterranean Sea is a wine-exporter

• Co-location rules
  – Association rules are generalized to data sets which are indexed by space; notion of transaction is replaced by neighbourhood
    • spatial items are discretized version of continuous variables
Example on trajectory data mining
A.I.S. data at 2009/04/03 12h00 (local time, Z+3)(Legouge, R.)
Data used as accident history source (Legouge, R.)
Space-time cube: trajectories as such
Space-time cube: densities and trajectories
Spatio-temporal patterns to look for

Spatio-temporal **hotspots**
- convergence of trajectories in both space and time

- many objects in near proximity at all times

Temporal **bridges**
- discrete temporal pattern at certain times

Temporal **towers**
5. Integrating human knowledge and data analysis

• In spatial data mining we combine
  – Computational, statistical and visual methods plus human interaction and human knowledge

• Example:
  1. Clustering
  2. Multivariate visualization of the contents of the clusters
  3. Human interpretation of the results
  4. Use of additional geographical data
  5. Final decision
Reading material and other references

• O´Sullivan,D., Unwin,D., Geographical Information Analysis.
• Shekhar,S., Chawla,S., Spatial Databases, A Tour, Chapter 7 "Introduction to spatial data mining".
• Longley,P., Rhind,D., Maquire,R., Goodchild,M., Geographic Information Systems and Science

• Figures taken from the doctoral theses/articles by:
  – Spatenkova,O.
  – Sunila,R.,
  – Krisp,J.
  – Zhe Zhang
  – Legouge,R.
Welcome to learn more on spatial analysis and spatial data mining on our courses!

Questions: now or later …

kirsi.virrantaus@aalto.fi