Spatial Clustering for Mining Knowledge in Support of Generalization Processes in GIS

Bin Jiang
University of Gävle, Sweden

- Introduction (structure, knowledge, spatial clustering)
- Related work (spatial data mining, Knowledge acquisition/knowledge discovery, generalization using spatial clustering)
- Spatial clustering: concepts and algorithms
- Generalization supported by spatial clustering: principles
  - Typification and selection rules,
  - Dendrogram as visual support
  - Generalization validation
- Two case studies (point pattern and street selection)
- Conclusion

- Generalization processes require maintaining the overall structures and patterns presented with the source map or database.
- Knowledge involved in generalization includes geometry, topological and semantic relationships, as well as procedural knowledge.
- Spatial clustering is a way of detecting groups of spatial objects, so that the objects in the same group are more similar than those in different groups.
- This paper aims to explore spatial clustering in support of generalization processes in GIS.
  - Cluster detection, visual support and validation

- Spatial data mining and knowledge discovery (Miller and Han 2001)
- Knowledge acquisition in generalization (Buttenfield and McMaster 1991)
- Generalization using spatial clustering
  - SOM-based clustering (Højholt 1995, Jiang and Harrie 2003)
  - Graph-based clustering (Jiang and Claramunt 2002, Anders 2003)

- Data is raw data, information is organized data, but "knowledge is power" (Francis Bacon 1561 – 1626)
- By knowledge it means…
  - hidden structures and patterns through spatial clustering.
  - parameter settings for clustering.
  - something of value for generalization emerged from interactive clustering processes.

- What is special for spatial clustering ?
- Two types of spatial clustering
  - Hierarchical, and
  - Non-hierarchical clustering
- *m* spatial objects with *n* dimensions:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & & & \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

- Similarity refers to how an object is similar to every other.
- A dendrogram that shows a tree-like hierarchy about cluster structure of objects.

- K-means algorithm:
  - First arbitrarily choose *k* centres and compute the distances from all the objects to the cluster centres, and then update the centres according to the new memberships of the objects. Repeat the procedure till the sum of distances from all objects in the individual clusters are minimised.
- Silhouette value:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

S(i)=1 suggests the object is rather centrally located

S(i)= 0 indicates that the object is on the edge of two neighbouring clusters

S(i)=–1 shows that the object is probably clustered into the wrong cluster

- A hierarchical clustering ends up with multiple partitions at the different levels of similarity
- Cophenetic correlation coefficient

$$\gamma = \frac{\sum_{i<j} (y_{ij} - \overline{y})(z_{ij} - \overline{z})}{\sqrt{\sum_{i<j} (y_{ij} - \overline{y})^2 \sum_{i<j} (z_{ij} - \overline{z})^2}}$$

  - The closer to unity the value is, the less distortion of the cluster tree is.

- How to support ? (principles)
- Typification ($X_i \Leftrightarrow c_j$)and selection ($\min_i\{| x_i - c_j |\}$) based on clustering processes
- Dendrogram as a visual support for multi-representation

$$(LOD_m,...LOD_2, LOD_1) \Leftarrow (p_m,...p_2, p_1)$$

- Cluster validity techniques for quality assessment of generalization

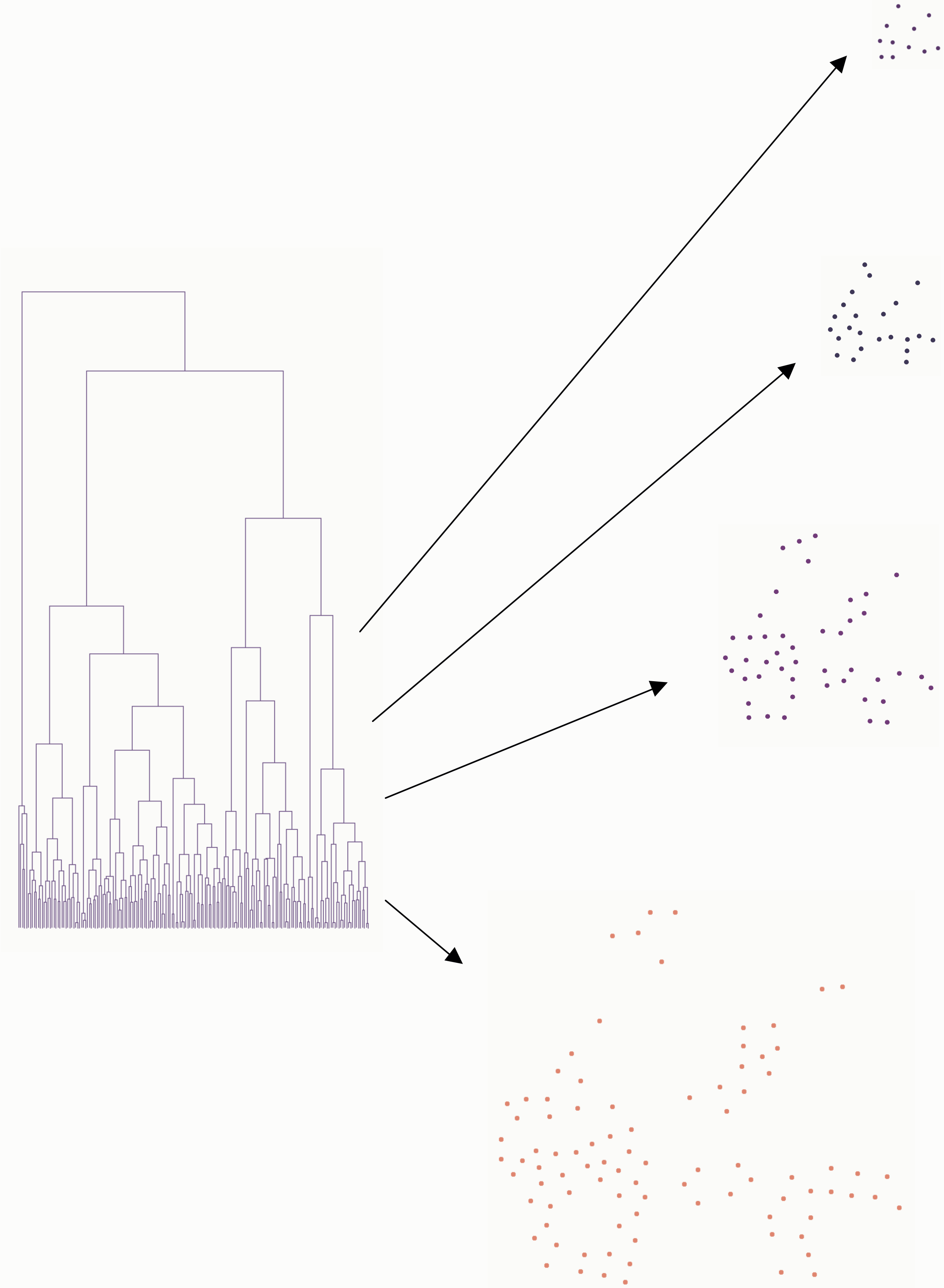
point pattern

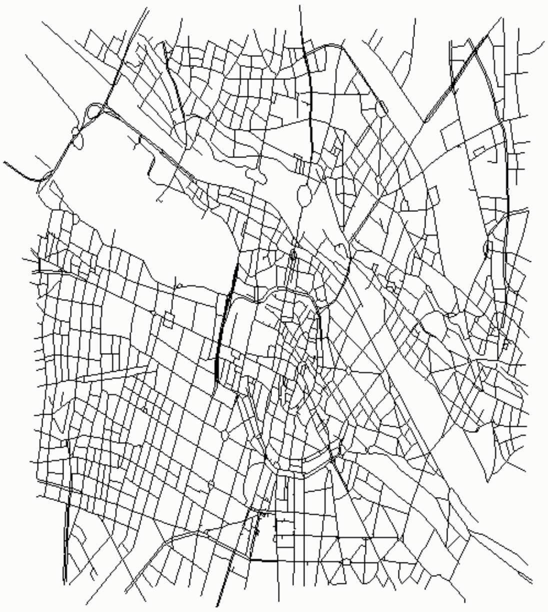
typification

selection

S-value for typ.

S-value for selec.





- To conclude …
- Spatial clustering can be important support for generalization processes in terms of pattern and knowledge discovery via
  - Cluster detection
  - Cluster validity, and
  - Visual support of dendrogram
- Future work aims to develop visual environments in support of knowledge discovery for generalization processes.