Spatially Constrained Clusters

Luc Anselin

http://spatial.uchicago.edu
basic principles
indirect solutions
skater
max-p
Basic Principles
• Problem

  grouping contiguous objects that are similar into new aggregate areal units

  tension between

  attribute similarity

  grouping of similar observations

  locational similarity

  group spatially contiguous observations only
• Terminology

regionalization (special case: redistricting)

spatially-constrained clustering

contiguity-constrained clustering

clustering under connectivity constraints

many different terms
• Multiple Objectives

  classical clustering

    maximize within-group similarity

    or, maximize between-group dissimilarity

  spatial similarity

    only contiguous objects in same group

  shape

    compactness
Solution Strategies (Duque et al. 2007)
• Classical Clustering with Updates

start with hierarchical clustering or k-means solution

split/combine clusters that are not contiguous

inefficient approach

number of cluster indeterminate
Multi-Objective Approach

introduce location \((x, y)\) as variables within the clustering routing

assign weights to similarity objective vs spatial objective

difficult to set weights
• Automatic Zoning

AZP

automatic zoning procedure (Openshaw and Rao)

heuristic

starts from random initial feasible solutions

optimization (NP-hard problem)
• **Graph-Based Approaches**

  represent the contiguity structure of the objects as a graph

  graph pruning

  e.g., using minimum spanning tree

  maximize internal similarity objective
Explicit Optimization

formulate as an integer programming problem

decision variables to allocate object $i$ to region $j$

formalize adjacency constraints

typically as a graph representation

several heuristics
Indirect Solutions
Classic Clustering with Updates
• Point of Departure - k Means Clusters

make any non-contiguous part of a cluster into a separate cluster

increases the number of clusters

fragmented solutions

move observations between clusters to achieve contiguity

keeps k the same

multiple solutions possible
k-means (k=4) solution

12 “contiguous” clusters
k-means (k=4) solution

4 contiguous clusters
six changes
<table>
<thead>
<tr>
<th>Method</th>
<th>Total SS</th>
<th>Within SS</th>
<th>Between SS</th>
<th>Ratio B/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>504</td>
<td>286.8</td>
<td>217.2</td>
<td>0.431</td>
</tr>
<tr>
<td>contiguous</td>
<td>504</td>
<td>314.8</td>
<td>189.2</td>
<td>0.375</td>
</tr>
<tr>
<td>k=12</td>
<td>504</td>
<td>237.4</td>
<td>266.6</td>
<td>0.529</td>
</tr>
</tbody>
</table>

cluster characteristics
Multi-Objective Optimization
Weighted Optimization

\[ w_1 \text{(attribute similarity)} + w_2 \text{(geometric centroids)} \]

\[ w_1 + w_2 = 1 \]

iterate until contiguity constraint is satisfied

bisection method

\[ w_2 \text{ is weight for centroids, } w_1 = 1 - w_2 \]

start with 0.0 and 1.0

then move to 0.50 - check contiguity

if contiguous, then to midpoint to the left of 0.50

if not contiguous, then to midpoint to the right of 0.50

etc… until contiguous with the highest bSS/tSS ratio
$w_2 = 0$

$bSS/tSS = 0.4338$

$w_2 = 1$

$bSS/tSS = 0.2461$
$w_2 = 0.50$

$bSS/tSS = 0.3474$

$w_2 = 0.25$

$bSS/tSS = 0.4166$
\[ w_2 = 0.375 \]
\[ bSS/tSS = 0.3680 \]

endpoint:

\[ w_2 = 0.4500 \]
\[ bSS/tSS = 0.3612 \]
ad hoc solution
ratio = 0.375

centroid solution
ratio = 0.361
skater
• **SKATER**

Spatial Kluster analysis by Tree Edge Removal


algorithm

construct minimum spanning tree from adjacency graph

prune the tree (cut edges) to achieve maximum internal homogeneity
Contiguity as a Graph

network connectivity based on adjacency between nodes (locations)

edge value reflects dissimilarity between nodes

\[ d(i,i') = d(x_i,x_{i'}) = \sum_p (x_{ip} - x_{i'p})^2 \]

objective is to minimize within-group dissimilarity (maximize between-group)
Queen contiguity network graph
Minimum Spanning Tree

connectivity graph $G = (V, L)$

$V$ vertices (nodes), $L$ edges

path

a sequence of nodes connected by edges

$v_1$ to $v_k$: $(v_1, v_2), \ldots, (v_{k-1}, v_k)$

spanning tree

tree with $n$ nodes of $G$

unique path connecting any two nodes

$n-1$ edges

minimum spanning tree

spanning tree that minimizes a cost function

minimize sum of dissimilarities over all nodes
**Minimum Spanning Tree Algorithm (Assuncao et al 2006)**

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First iteration</strong></td>
<td>Set $T_1 = (V_1, L_1)$, where $V_1 = {v_1}$ and $L_1 = \emptyset$. Find the edge of lowest cost $(l_{a}, l_{f})$. Step 3: $T_2 \Rightarrow V_2 = {v_1, v_3}$ and $L_2 = {l_{a}}$. Step 4: Repeat Step 2.</td>
</tr>
<tr>
<td><strong>Second iteration</strong></td>
<td>Find the edge of lowest cost $(l_{b}, l_{c}, l_{d}, l_{e}, l_{f})$. Set $T_3 \Rightarrow V_3 = {v_1, v_3, v_5}$ and $L_3 = {l_{a}, l_{b}}$.</td>
</tr>
<tr>
<td><strong>Third iteration</strong></td>
<td>Find the edge of lowest cost $(l_{c}, l_{d}, l_{e}, l_{f})$. Set $T_4 \Rightarrow V_4 = {v_1, v_3, v_4, v_5}$ and $L_3 = {l_{a}, l_{b}, l_{c}}$.</td>
</tr>
<tr>
<td><strong>Final Iteration</strong></td>
<td>$V_n = V$.</td>
</tr>
</tbody>
</table>

Figure 2. Construction of the minimum spanning tree.
• Tree Pruning

finding spatially contiguous clusters as a tree partitioning problem

to obtain $k$ regions, $k-1$ edges need to be removed

removal of edges results in sub-trees = cluster

hierarchical approach

minimize within-cluster sum of squares

cut where $\max F(T) - [F(T_a) + F(T_b)]$

with $F(T)$ as the within SS for tree $T$
Iteration 0: $G^* = \text{MST}$. We select the edge which has the largest objective function. Cut out this edge leaving two trees ($T_1$ and $T_2$).

Iteration 1: $G^* = (T_1, T_2)$. We compare the highest objective functions for $T_1$ and $T_2$. We split the tree $T_1$ since $f_i(S^{T_2}_*) \leq f_i(S^{T_1}_*)$

Iteration 2: $G^* = (T_2, T_3, T_4)$. We compare the highest objective functions for $T_2$, $T_3$ and $T_4$. We split the tree $T_3$ since $f_i(S^{T_4}_*) \leq f_i(S^{T_3}_*) \leq f_i(S^{T_4}_*)$

Figure 3. Partitioning of the MST.

skater - pruning the MST (Assuncao et al 2006)
skater clusters k=4
SSw = 344.9
SSb = 159.1
SSb/SSt = 0.316

skater clusters k=4
skater clusters k=6
SSw = 292.6

SSb = 211.4

SSb/SSt = 0.420

skater clusters k=6
• Issues

constrains solution space

only cuts in MST and subsets of MST

local optima

doesn’t scale well
max-p
- Selecting k

  ad hoc rules

  plot ratio between SS / total SS by k

  plot ratio within SS / total SS by k

  find “elbow” (similar to scree plot for PCA)
ratio between SS / total SS by number of clusters
k-means
ratio within SS / total SS by number of clusters

k-means
Max-p Regions Problem

aggregation of n areas into an unknown maximum number (p) of homogenous regions

each region satisfies a minimum threshold on a spatially extensive variable (e.g., population, area)

number of regions is endogenous

data dictate shape of regions

contiguity enforced, but not compactness
Problem Formulation

Parameters:

\[ i, I = \text{Index and set of areas, } I = \{1, \ldots, n\}; \]
\[ k = \text{index of potential regions, } k = \{1, \ldots, n\}; \]
\[ c = \text{index of contiguity order, } c = \{0, \ldots, q\}, \text{ with } q = (n - 1); \]
\[ w_{ij} = \begin{cases} 1, & \text{if areas } i \text{ and } j \text{ share a border, with } i, j \in I \text{ and } i \neq j \\ 0, & \text{otherwise}; \end{cases} \]
\[ N_i = \{j \mid w_{ij} = 1\}, \text{ the set of areas that are adjacent to area } i; \]
\[ d_{ij} = \text{dissimilarity relationships between areas} \]
\[ i \text{ and } j, \text{ with } i, j \in I \text{ and } i < j; \]
\[ h = 1 + \lfloor \log(\sum_i \sum_{j>i} d_{ij}) \rfloor, \text{ which is the number of digits of the} \]
\[ \text{floor function of } \sum_i \sum_{j>i} d_{ij}, \text{ with } i, j \in I; \]
\[ l_i = \text{spatially extensive attribute value of area } i, \text{ with } i \in I; \]
\[ \text{threshold} = \text{minimum value for attribute } l \text{ at regional scale}. \]

Decision variables:

\[ t_{ij} = \begin{cases} 1, & \text{if areas } i \text{ and } j \text{ belong to the same region } k, \text{ with } i < j \\ 0, & \text{otherwise}; \end{cases} \]
\[ x_{i}^{kc} = \begin{cases} 1, & \text{if areas } i \text{ is assigned to region } k \text{ in order } c \\ 0, & \text{otherwise}. \end{cases} \]
Problem Formulation (2)

\[
\begin{align*}
\text{Minimize:} & \quad \quad Z = \left( \sum_{k=1}^{n} \sum_{i=1}^{n} x_{ik}^{k0} \right) \times 10^k + \sum_{i} \sum_{j, j > i} d_{ij} t_{ij}. \\
\text{Subject to:} & \quad \quad \sum_{i=1}^{n} x_{ik}^{k0} \leq 1 \quad \forall k = 1, \ldots, n; \\
& \quad \quad \sum_{k=1}^{n} \sum_{c=0}^{q} x_{ik}^{kc} = 1 \quad \forall i = 1, \ldots, n; \\
& \quad \quad x_{ik}^{kc} \leq \sum_{j \in N_i} x_{jk}^{k(c-1)} \quad \forall i = 1, \ldots, n; \forall k = 1, \ldots, n; \forall c = 1, \ldots, q; \\
& \quad \quad \sum_{i=1}^{n} \sum_{c=0}^{q} x_{ik}^{kc} l_i \geq \text{threshold} \times \sum_{i=1}^{n} x_{ik}^{k0} \quad \forall k = 1, \ldots, n; \\
& \quad \quad t_{ij} \geq \sum_{c=0}^{q} x_{ik}^{kc} + \sum_{c=0}^{q} x_{jk}^{kc} - 1 \quad \forall i, j = 1, \ldots, n|i < j; \forall k = 1, \ldots, n; \\
& \quad \quad x_{ik}^{kc} \in \{0, 1\} \quad \forall i = 1, \ldots, n; \forall k = 1, \ldots, n; \forall c = 0, \ldots, q; \\
& \quad \quad t_{ij} \in \{0, 1\} \quad \forall i, j = 1, \ldots, n|i < j.
\end{align*}
\]
Logic of Objective Function

first term controls the number of regions

second term controls pairwise dissimilarities

first term dominates (scaling factor)

solution with higher value of $p$ will always be preferred over lower $p$ in terms of dissimilarity

for same value of $p$, solutions with lower heterogeneity are preferred

avoids comparing heterogeneity between regions for different $p$
• **Logic of Constraints**

Each region starts with a root area $x_{i}^{k0}$ to which other areas are added that are contiguous.

In each region, there can only be one area of a given order of contiguity to the root area.

The spatially extensive variable summed over all areas in the region must meet the threshold.
Solution Strategies

mixed integer programming

exact solution impractical

heuristics

construction phase: set of feasible solutions

local search phase: iterative improvements

simulated annealing

tabu search

greedy algorithm
population threshold 10%
\[p = 8\]
\[\frac{bSS}{tSS} = 0.525\]

population threshold 20%
\[p = 4\]
\[\frac{bSS}{tSS} = 0.375\]

max p results
ad hoc — 0.375

centroids — 0.361

skater — 0.316

k-means 0.431

max p — 0.375
Summary

trade-off attribute similarity and locational similarity is complex

no “best” approach

no mechanical application of one approach

sensitivity analysis is critical