Clustering Components of PySAL

Sergio Rey$^{1,2}$  Juan Carlos Duque$^1$  Luc Anselin $^{2,3}$

1 Regional Analysis Laboratory (REGAL)
Department of Geography
San Diego State University

2 Regional Economics Application Laboratory (REAL)
University of Illinois Urbana Champaign

3 Department of Geography
University of Illinois Urbana Champaign

Regional Science Association International
Las Vegas, Nevada
November 10-12, 2005
Outline

1 PySAL
   - Origins
   - Objectives

2 Clustering Components
   - Definition
   - Methods

3 Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4 Future Directions
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date…
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions

Sergio Rey, Juan Carlos Duque, Luc Anselin
Outline

1 PySAL
   - Origins
   - Objectives

2 Clustering Components
   - Definition
   - Methods

3 Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4 Future Directions
1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Collaborative Project

- OpenSpace (UIUC)
- STARS (SDSU)
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions

Sergio Rey, Juan Carlos Duque, Luc Anselin
Objectives

Collaborative Project

- OpenSpace (UIUC)
- STARS (SDSU)
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Contiguity Constrained Clustering

Aggregation of $N$ areas to $M$ regions ($M < N$), such that:

1. Each area belong to only one region.
2. The areas assigned to a region must be geographically connected.

$N = 47$
$M = 6$
Outline

1 PySAL
   - Origins
   - Objectives

2 Clustering Components
   - Definition
   - Methods

3 Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4 Future Directions
Methods

Geographic data reduction
  - Supervised
    - Without contiguity constrain
    - In two stages
  - Unsupervised
    - With contiguity constrain
    - Inclusion of geographical information in the set of classification variables
    - Use of additional instruments
    - Symmetric relations
    - Non-symmetric relations
    - Hierarchical algorithms
    - Partitioning algorithms
    - Mathematical programming
    - Iterative algorithms / heuristics
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARiSeL
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions

Sergio Rey, Juan Carlos Duque, Luc Anselin
Clustering in PySAL
RSAI 2005
Spatial Aggregation Module

Multivariate euclidean distance
Mult. euclid. distance with spatial component
Distance functions for panel data
Contiguity matrix simulator
Contiguity matrix format translator (GAL, links, W)
Contiguity constrain checker
Random aggregation with contiguity constrain
Solutions overlapping tool (intersections)
Intra cluster distances
Distance to centroids
Theil index
Homogeneity
Compactness
Equality

Spatial Aggregation

Distance
Contiguity
Miscellany

Models
Two-stages models
k-means in two stages
With geographical attributes (induced contiguity)
With contiguity matrix
Greedy+TABU
AZP-TABU
MST

Conventional cluster with weighted distance component

xy, var/dist
var/dist, W, [s]
var/dist, W, p
var/linkcost, W

i: optional input
var: variables
dist: distances
W: connectivity
xy: coordinates
s: initial seeds
p: initial feasible partition
1. **PySAL**
   - Origins
   - Objectives

2. **Clustering Components**
   - Definition
   - Methods

3. **Aggregation Models in PySAL**
   - Forward Planning
   - **Up to date...**
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISeL
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. **Future Directions**
Menu

Clustering in PySAL
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISeL
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Parameters:
\( W \)  Information about contiguity between areas to be aggregated.
\( V \)  Matrix \((n \cdot p)\) with \(n\) observations and \(p\) variables.
\( m \)  Number of regions to be designed.

---

**Procedure 1.1: K-means Two Stages** \((W, V, m)\)

**comment:** Aggregate \(n\) small zones into \(m\) regions, \(m < n\).

\[ \Delta^* \leftarrow \infty \]

\[ K \leftarrow m \]

while \( Flag = 0 \)

\[
1: \phi \leftarrow \text{Regional configuration obtained by applying K-means to design} \ K \ \text{clusters} \\
2: m' \leftarrow \text{The number of contiguous regions in} \ \phi \\
3: \Delta \leftarrow \text{abs}(m - m')
\]

do

if \( \Delta < \Delta^* \)

then

\[
\begin{align*}
\Delta^* &\leftarrow \Delta \\
K &\leftarrow m - 1
\end{align*}
\]

else if \( \Delta \geq \Delta^* \)

then

\[
\begin{align*}
K &\leftarrow K + 1 \\
Flag &\leftarrow 1
\end{align*}
\]

4: \( \phi' \leftarrow \text{Regional configuration obtained by applying K-means to design} \ K' \ \text{clusters} \\
return \ (\phi')
K-means two stages

Two-Stage K-Means Clustering

- kmeans variable(s)
  - income
  - income_L
  - pcr
  - pcr_L
  - x
  - y

- Time Periods
  - 1995
  - 1996
  - 1997
  - 1998
  - 1999
  - 2000

Number
Number of Clusters

Choose a Matrix
States48

User Defined Variable Name
New Variable Name

Graphs Options
Map All None

OK Cancel
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Including Centroids Coordinates

```
Two-Stage K-Means Clustering

- kmeans variable(s):
  - income
  - income_L
  - perc
  - perc_L
  - x
  - y

- Time Periods:
  - 1995
  - 1996
  - 1997
  - 1998
  - 1999
  - 2000

Number
Number of Clusters

Choose a Matrix:
- States48

User Defined Variable Name
New Variable Name

Graphs Options:
- Map
- All
- None

OK  Cancel
```
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
     - Automatic Regionalization with Initial Seeds Location: ARISeL
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Parameters:
\( W \)  Information about contiguity between areas to be aggregated.
\( V \)  Matrix \((n \cdot p)\) with \( n \) observations and \( p \) variables.
\( m \)  Number of regions to be designed.
\( R \)  Length of prohibition.
\( C \)  Number of times the solution is equal to the aspirational criterion.
\( t \)  Maximum running time.

---

**Procedure 1.2: AZP-tabu**\( (W, V, m, R, C, t) \)

**comment:** Aggregate \( n \) small zones into \( m \) regions, \( m < n \).

\( \phi \leftarrow \) Start with a random zoning system of \( n \) small zones into \( m \) regions, \( m < n \). *NOTE: In our code, the initial partition comes from running the first part of ARISel, with \( S=1 \)*

repeat

1: Find the global best move that is not prohibited or tabu.
2: Make this move if it is an improvement or equivalent in value, else:
3: If no improving move can be made, then see if a tabu move can be made which improves on the current local best (termed an aspiration move), else:
4: If there is no improving and no aspirational move, then make the best move even if it is nonimproving (that is, results in a worse value of the objective function).
5: Tabu the reverse move for \( R \) iterations.

until Convergence criterion, \( C \) or \( T \), is satisfied

return \( \phi \)
AZP-tabu

Clustering in PySAL
Comparison

K-means two stages = 59.346
ARISeL = 45.599
AZP-Tabu = 48.088
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - **Automatic Regionalization with Initial Seeds Location: ARISel**
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions
Parameters:

- $W$: Information about contiguity between areas to be aggregated.
- $V$: Matrix $(n \cdot p)$ with $n$ observations and $p$ variables.
- $m$: Number of regions to be designed.
- $S$: Number of seeds relocation.
- $R$: Length of prohibition.
- $C$: Number of times the solution is equal to the aspirational criterion.
- $t$: Maximum running time.
Procedure 1.3: ARIS\_EL(W, V, m, S, R, C, T)

**comment:** Aggregate $n$ small zones into $m$ regions, $m < n$.

$\varphi \leftarrow \varphi$

$OF(\varphi) \leftarrow \infty$

**FIRST PART** (looking for the best initial partition)

1: List of seed: Select, at random, $m$ areas from the $n$ areas to be grouped and assign them to different regions.

for $l \leftarrow 1$ to $S$

repeat

for $growingRegion \leftarrow 1$ to $m$

2: Identify the bordering unassigned areas ($bua$).

3: Estimate the average dissimilarity between areas in $growingRegion$ and each $bua$.

4: Assignment: Select the $bua$ with smaller average dissimilarity and append it to the bordering $growingRegion$

5: Update objective function

until All the areas have been assigned: we have a newPartition

if $OF(newPartition) < OF(\varphi)$

then

6: $\varphi \leftarrow newPartition$

7: $OF \leftarrow OF(newPartition)$

Continue...
8: New potential seed: Calculate the increment pattern of the objective function along the assignation process and identify the area that causes the biggest increment. This area is a “new potential seed” (nps).

9: Candidate areas to leave the seed list (l₁ and l₂): Identify the closest regions in terms of dissimilarity and its respective seeds l₁ and l₂.

10: Update the list of seeds: replace l₁ or l₂ with nps.

11: S ← S + 1.

SECOND PART (local search)

φ ← AZP-TABU(W,V,m,R,C,t) : NOTE: Use φ as initial feasible solution

return (φ)
ARISEL Clustering

clustering variable(s)
- pc
- income
- income_L
- pcr
- pc_L
- y

Time Periods
- 1995
- 1996
- 1997
- 1998
- 2000

Choose a Matrix
- States18
- Matrix

Number
- Number of Clusters
- Number of Reallocations
- Length of Prohibition
- Convergence Criterion
- Convergence Time Restriction

User Defined Variable Name
New Variable Name

Graphs Options
- Map
- All
- None

ARISEL Clustering Results

ARISEL_7_2000

States-18 Matrix

User Defined Variable Name
New Variable Name

ARISEL Clustering Results

ARISEL_7_2000

States-18 Matrix

User Defined Variable Name
New Variable Name

ARISEL Clustering Results

ARISEL_7_2000

States-18 Matrix

User Defined Variable Name
New Variable Name
Outline

1. PySAL
   - Origins
   - Objectives

2. Clustering Components
   - Definition
   - Methods

3. Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - **Minimum Spanning Tree**
   - Integrating Local Indicators of Spatial Association (LISA)

4. Future Directions

Sergio Rey, Juan Carlos Duque, Luc Anselin

Clustering in PySAL

RSAI 2005 30 / 36
Outline

1 PySAL
   - Origins
   - Objectives

2 Clustering Components
   - Definition
   - Methods

3 Aggregation Models in PySAL
   - Forward Planning
   - Up to date...
   - K-means two stages
   - Including Centroids Coordinates
   - AZP-tabu
   - Automatic Regionalization with Initial Seeds Location: ARISel
   - Minimum Spanning Tree
   - Integrating Local Indicators of Spatial Association (LISA)

4 Future Directions
Integrating Local Indicators of Spatial Association (LISA)

\[ y_l_i = f(w_i, y) \]  \hspace{1cm} (1)

Integrating into regionalization algorithms

- **Kmeans** Select \( k \) significant LISAs to serve as initial seeds.
- **Kmeans** Extend attribute set to include local statistics
- **Kmeans** Combine: LISA seeds and LISA attributes
Future Directions

Regionalization
Extensions to what we have done here.

PySAL Clustering
Other aspects of Clustering in PySAL