STICC: A multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity

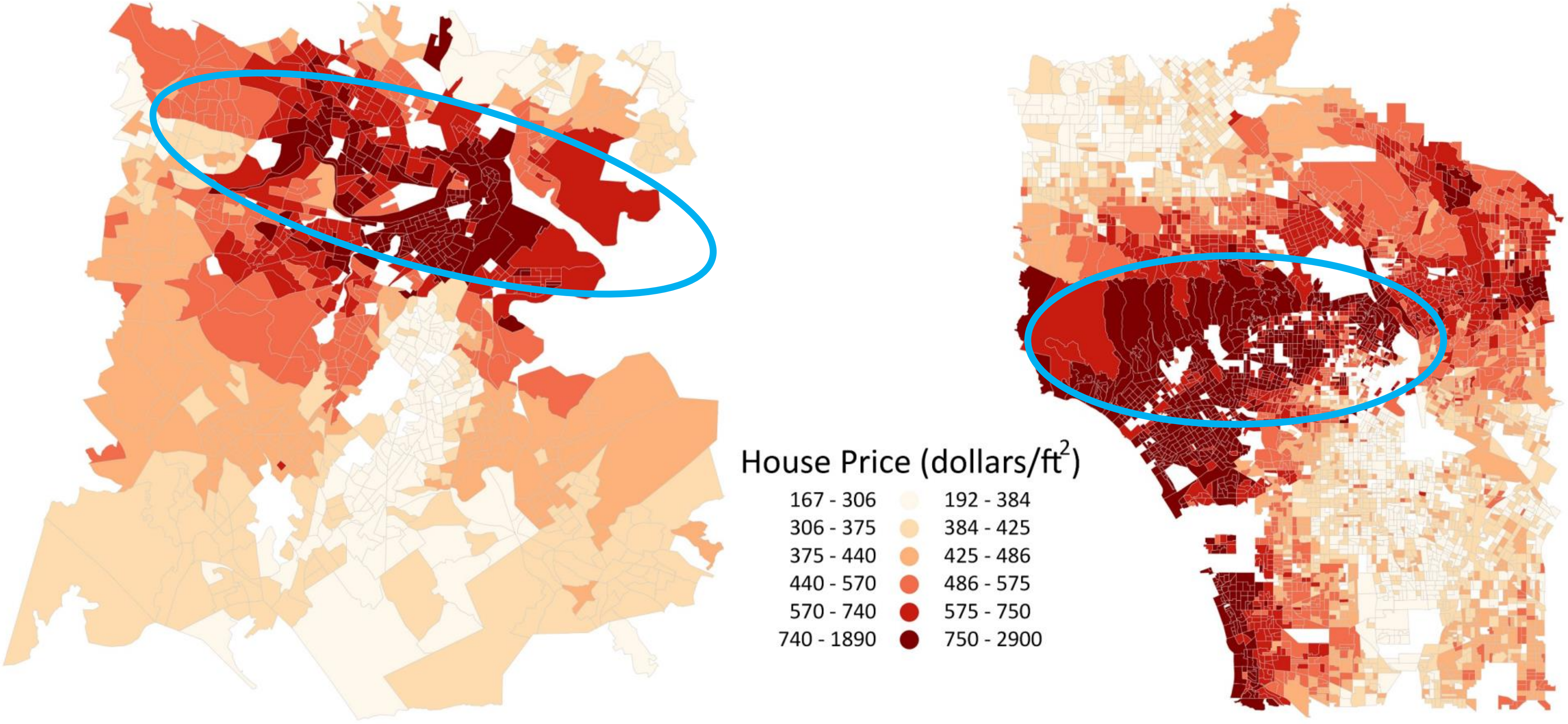
Yuhao Kang

Geospatial Data Science Lab, Department of Geography, University of Wisconsin-Madison

Kang, Y., Wu, K., Gao, S., Ng, I., Rao, J., Ye, S., Zhang, F. and Fei, T., 2022. STICC: A multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical Information Science*

Repeated Geographic Pattern

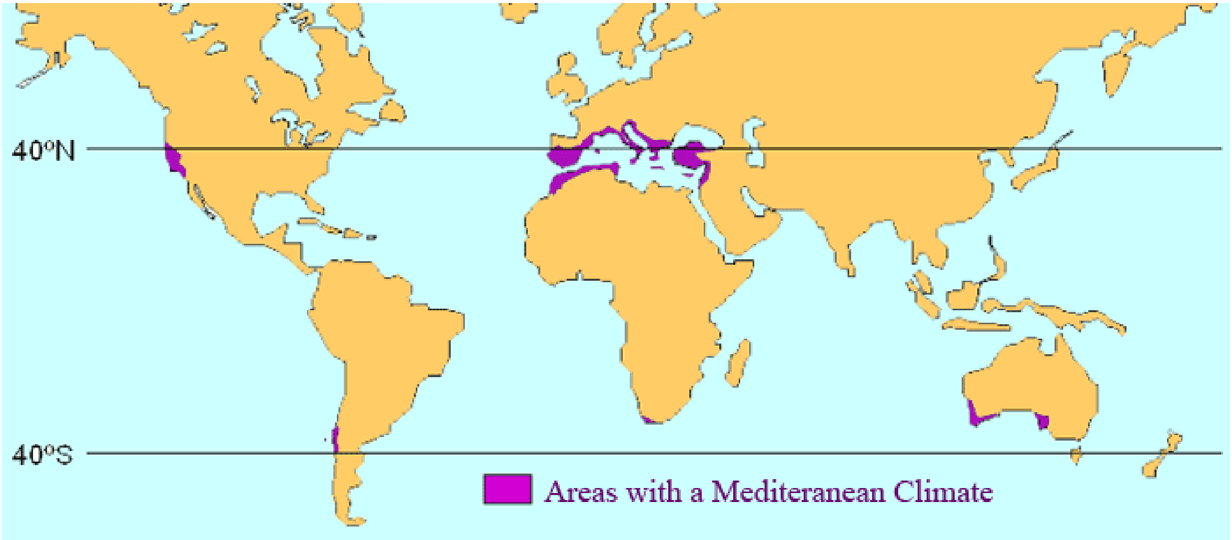
Spatial distribution of similar places:
1. Nearby places share similar characteristics (The first law of Geography)



Nearby neighborhoods may have similar house prices

Repeated Geographic Pattern

Spatial distribution of similar places:
2. Places located in different areas may have similar attributes



Italy and California, US, have the same Mediterranean climate type

Airports in different regions are both transportation hubs

Repeated Geographic Pattern Discovery (RGPD)

Definition

Finding out *repeated* groups of similar places across space and maintaining the *spatial contiguity* of geographic patterns within each subcluster

Repeated Geographic Pattern Discovery (RGPD)

Definition

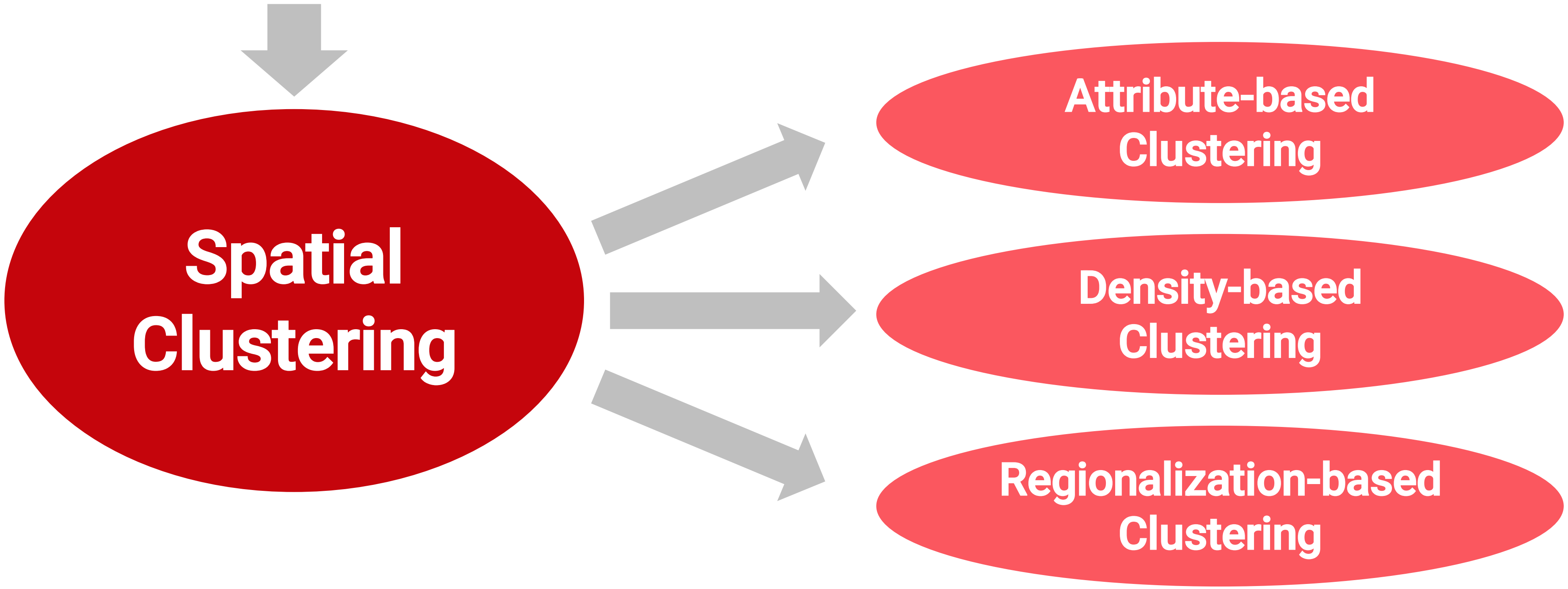
Finding out *repeated* groups of similar places across space and maintaining the *spatial contiguity* of geographic patterns within each subcluster

Spatial Clustering

Repeated Geographic Pattern Discovery (RGPD)

Definition

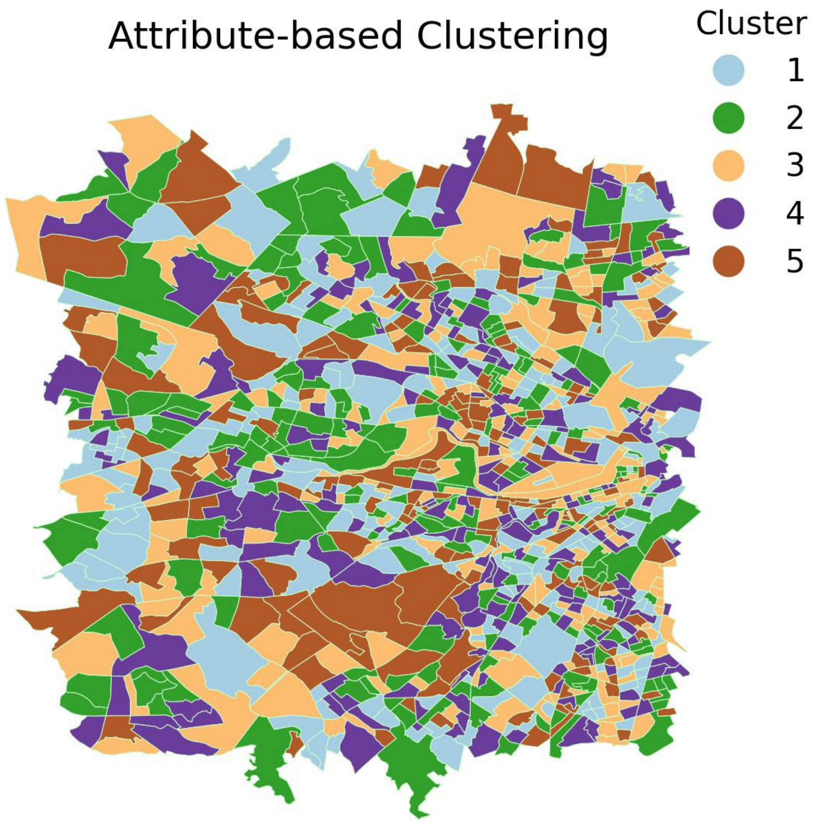
Finding out *repeated* groups of similar places across space and maintaining the *spatial contiguity* of geographic patterns within each subcluster

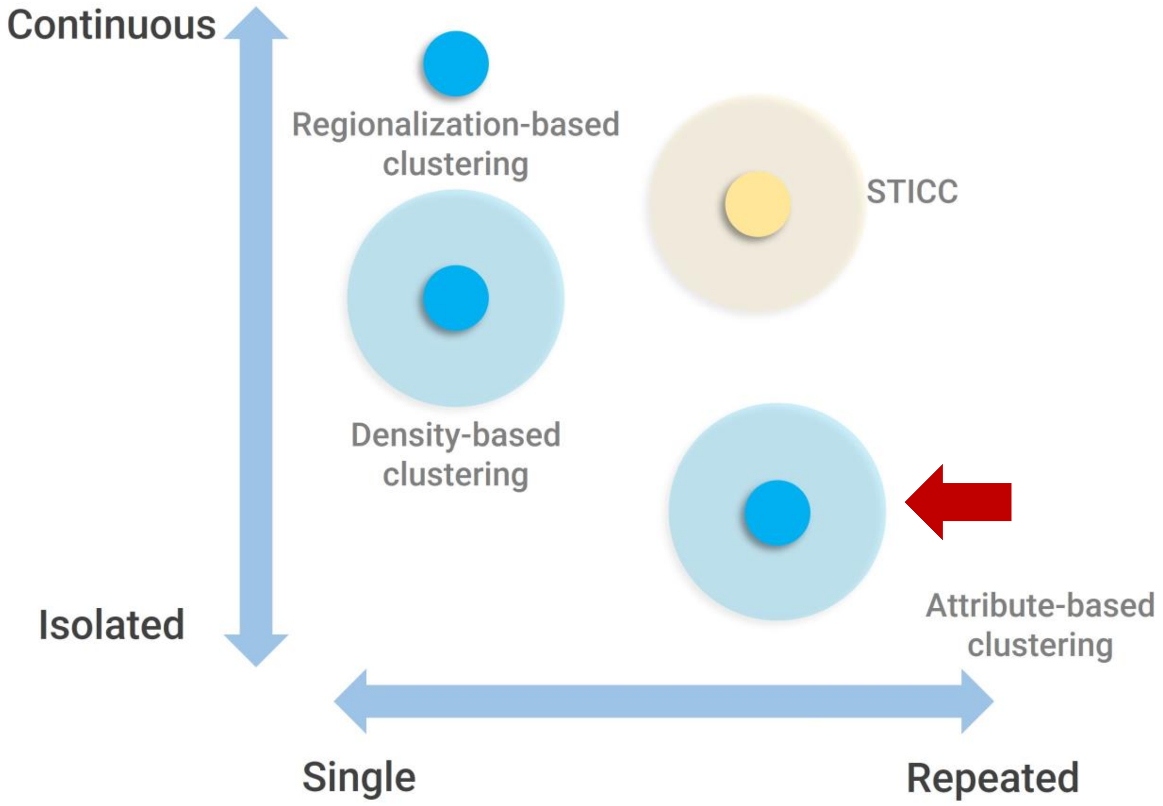


Attribute-based Clustering

Examples of attribute-based clustering:

K-Means, BIRCH, CURE, and SOM (self-organized map)

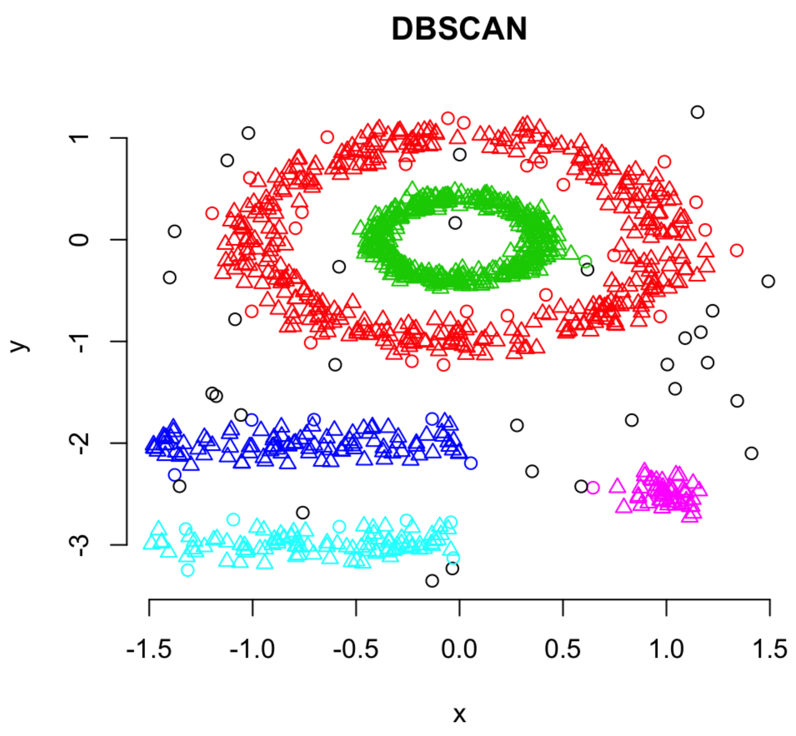


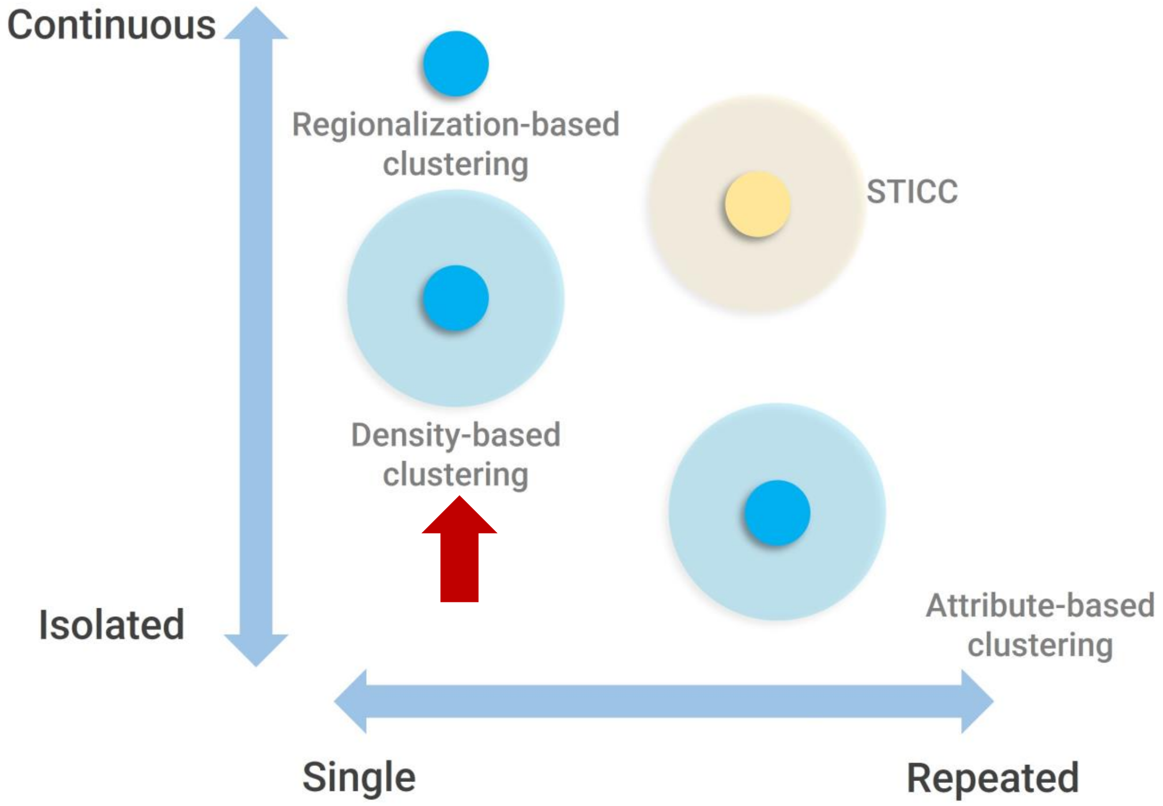


Density-based Clustering

Examples of density-based clustering:

DBSCAN, OPTICS, ENCLUE, ADCN

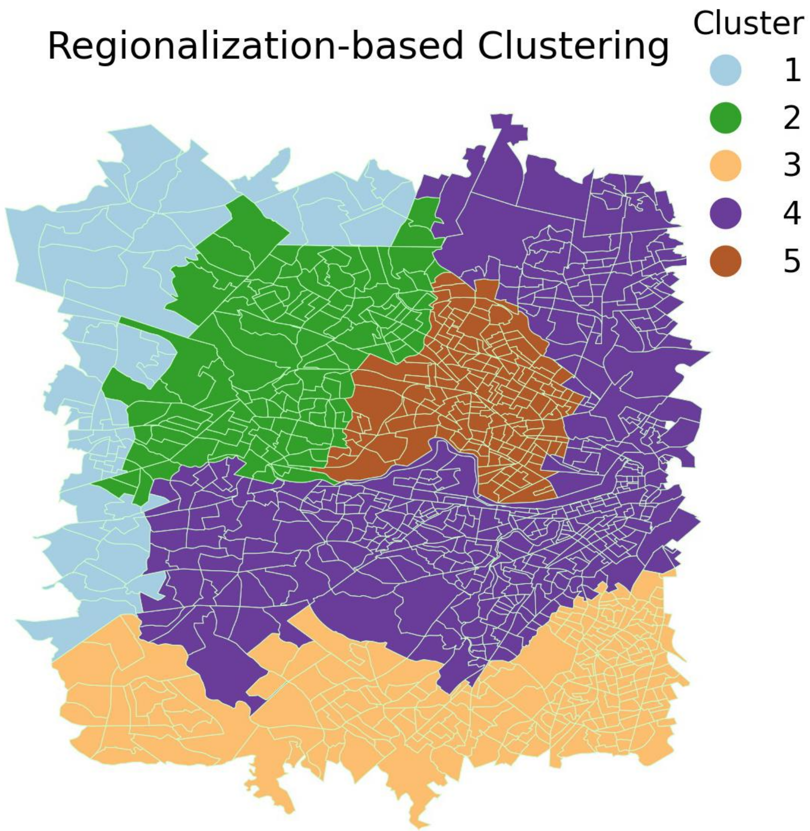


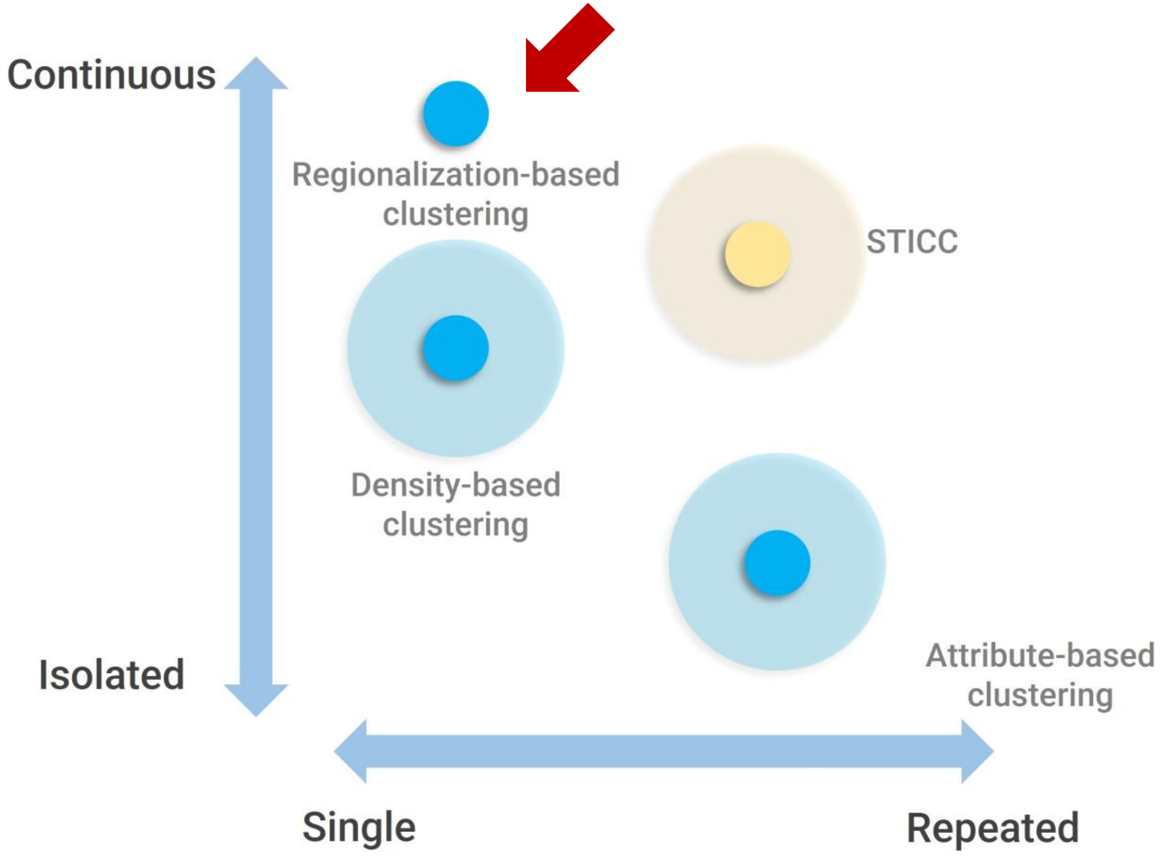


Regionalization-based Clustering

p-regions problem: the aggregation of *n* small areas into *p* geographically connected regions

Examples of regionalization-based clustering: SKATER, AUTOCLUST

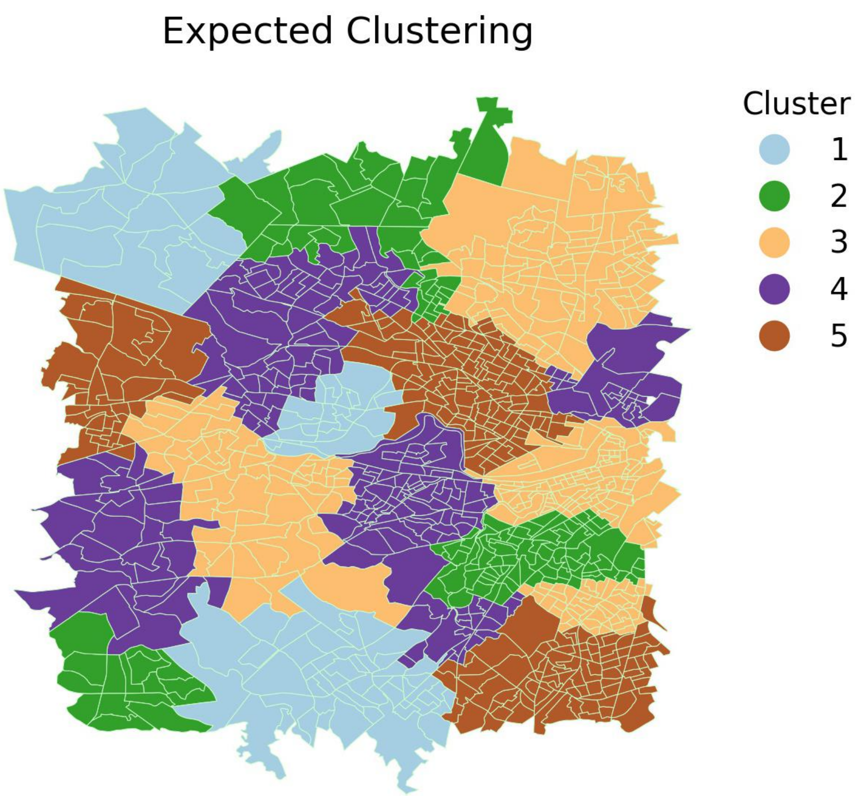


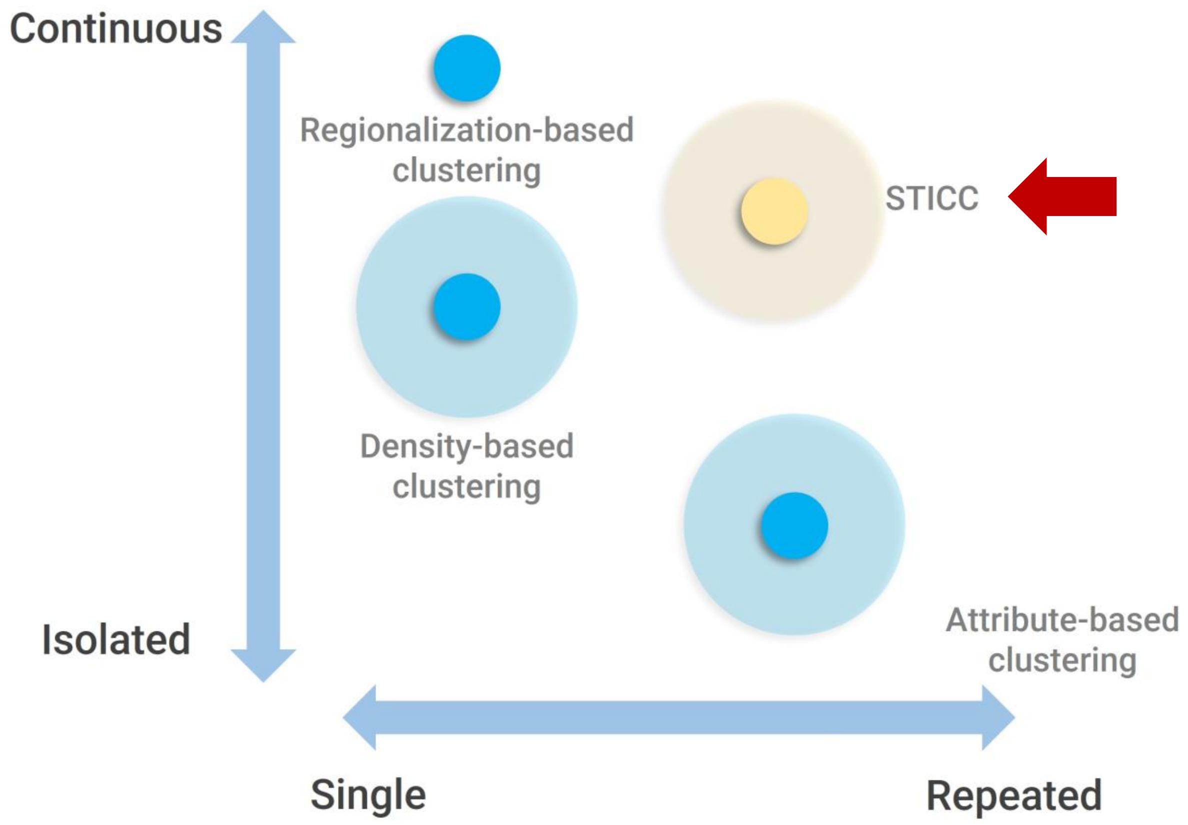


Spatial Toeplitz Inverse Covariance-based Clustering

Objective

Find *repeated* geographic patterns and maintain *spatial contiguity* simultaneously





Spatial Toeplitz Inverse Covariance-based Clustering

Features

1. Markov random field (MRF) for modeling partial correlations within each *subregion*
2. A *spatial consistency* strategy to encourage the nearby geographic objects to belong to the same cluster

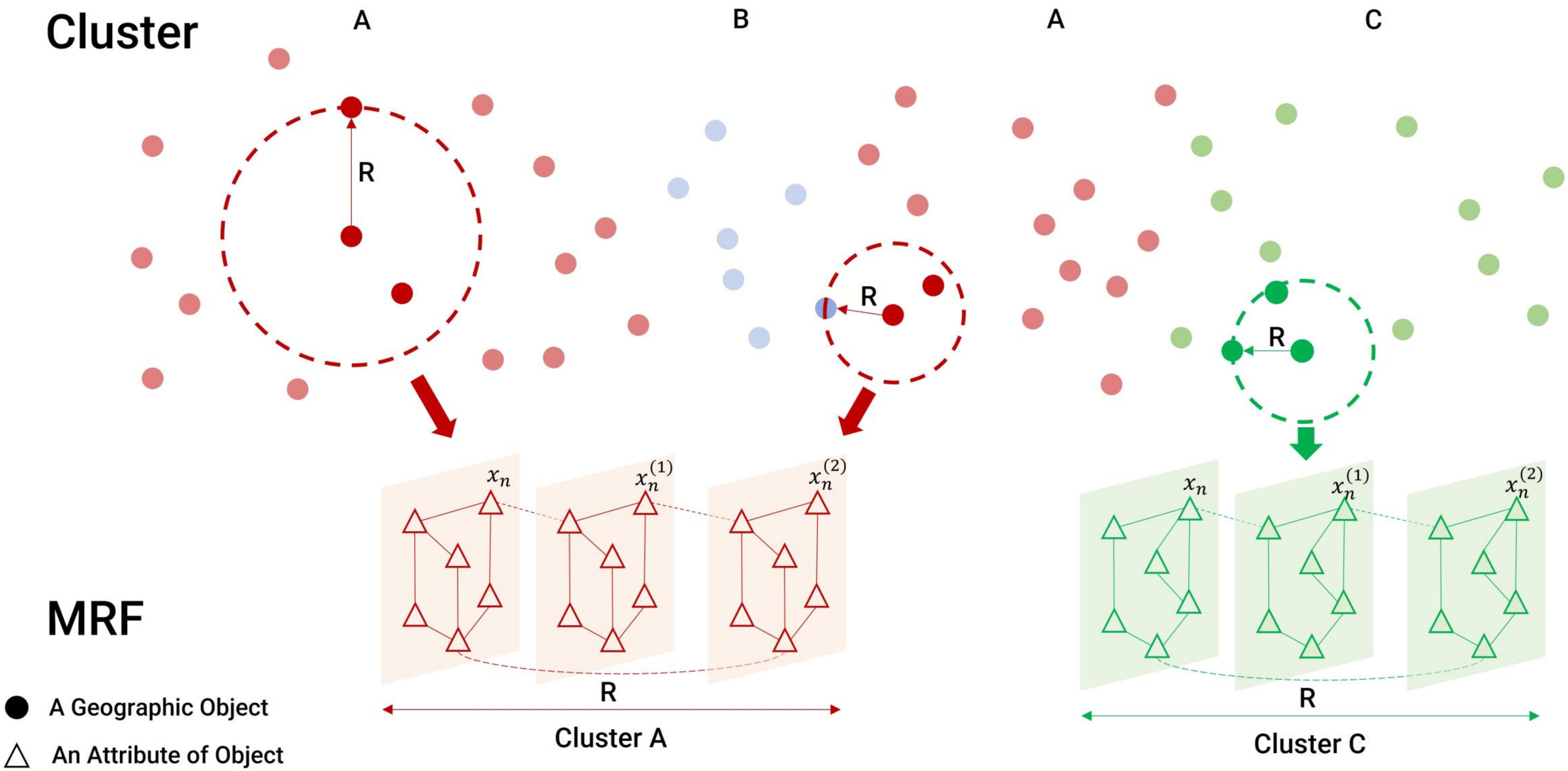
Motivation

TICC for multivariate time series data clustering

Ordered sequence in time series data but no orders in spatial data

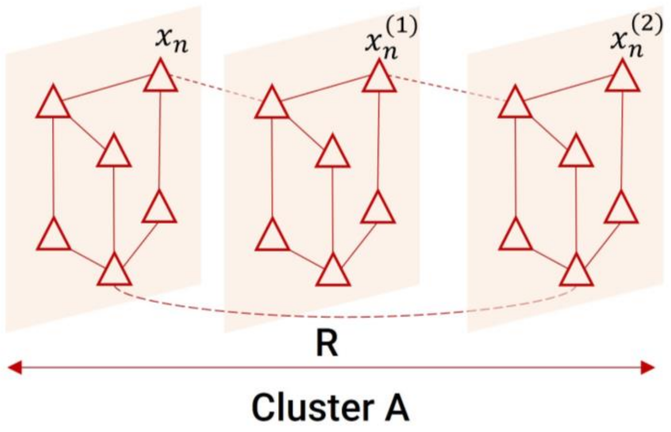
Hallac, D., Vare, S., Boyd, S. and Leskovec, J., 2017, August. Toeplitz inverse covariance-based clustering of multivariate time series data. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 215-223).

STICC Ideas



Keywords Subregion, radius (*R*), Markov Random Field

Spatial Toeplitz Matrix



$$\Theta_k = \begin{bmatrix} A_k^{(0)} & (A_k^{(1)})^T & (A_k^{(2)})^T & \cdots & \cdots & (A_k^{(R-1)})^T \\ A_k^{(1)} & A_k^{(0)} & (A_k^{(1)})^T & \ddots & & \vdots \\ A_k^{(2)} & A_k^{(1)} & A_k^{(0)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & (A_k^{(1)})^T & (A_k^{(2)})^T \\ \vdots & & \ddots & A_k^{(1)} & A_k^{(0)} & (A_k^{(1)})^T \\ A_k^{(R-1)} & \cdots & \cdots & A_k^{(2)} & A_k^{(1)} & A_k^{(0)} \end{bmatrix}$$

The inverse covariance matrices follows the Toeplitz form

Overall STICC Optimization Problem

Objective

$$\min_{\Theta \in \mathcal{T}, P} \sum_{k=1}^{K} \left[\sum_{X_n \in P_k} \left(\overbrace{-\mathcal{L}(\Theta_k; X_n)}^{\text{log likelihood}} + \overbrace{\beta \, \mathbb{1}\{X_n^{(1)} \notin P_k\}}^{\text{spatial consistency}} \right) + \overbrace{\|\lambda \odot \Theta_k\|_{1,\text{off}}}^{\text{sparsity}} \right]$$

Overall STICC Optimization Problem

Objective

$$\min_{\Theta \in \mathcal{T}, P} \sum_{k=1}^{K} \left[\sum_{X_n \in P_k} \left(\overbrace{\boxed{-\mathcal{L}(\Theta_k; X_n)}}^{\text{log likelihood}} + \overbrace{\beta \, \mathbb{1}\{X_n^{(1)} \notin P_k\}}^{\text{spatial consistency}} \right) + \overbrace{\|\lambda \odot \Theta_k\|_{1,\text{off}}}^{\text{sparsity}} \right]$$

$$\overbrace{-\mathcal{L}(\Theta_k; X_n)}^{\text{log likelihood}}$$

The negative log likelihood that the subregion X_n belongs to the *k*th cluster

Overall STICC Optimization Problem

Objective

$$\min_{\Theta \in \mathcal{T}, P} \sum_{k=1}^{K} \left[\sum_{X_n \in P_k} \left(\boxed{\overbrace{-\mathcal{L}(\Theta_k; X_n)}^{\text{log likelihood}}} + \boxed{\overbrace{\beta \, \mathbb{1}\{X_n^{(1)} \notin P_k\}}^{\text{spatial consistency}}} \right) + \overbrace{\|\lambda \odot \Theta_k\|_{1,\text{off}}}^{\text{sparsity}} \right]$$

$$\overbrace{-\mathcal{L}(\Theta_k; X_n)}^{\text{log likelihood}}$$

The negative log likelihood that the subregion X_n belongs to the *k*th cluster

$$\overbrace{\beta \, \mathbb{1}\{X_n^{(1)} \notin P_k\}}^{\text{spatial consistency}}$$

$$\mathbb{1}\{X_n^{(1)} \notin P_k\} = \begin{cases} 0, & \text{if } X_n, X_n^{(1)} \in P_k, \\ 1, & \text{otherwise.} \end{cases}$$

If the subregion X_n and its nearest neighbor $X_n^{(1)}$ belong to the same cluster then no cost, and vice versa

Overall Steps for STICC

Algorithm 1 Overall steps for STICC

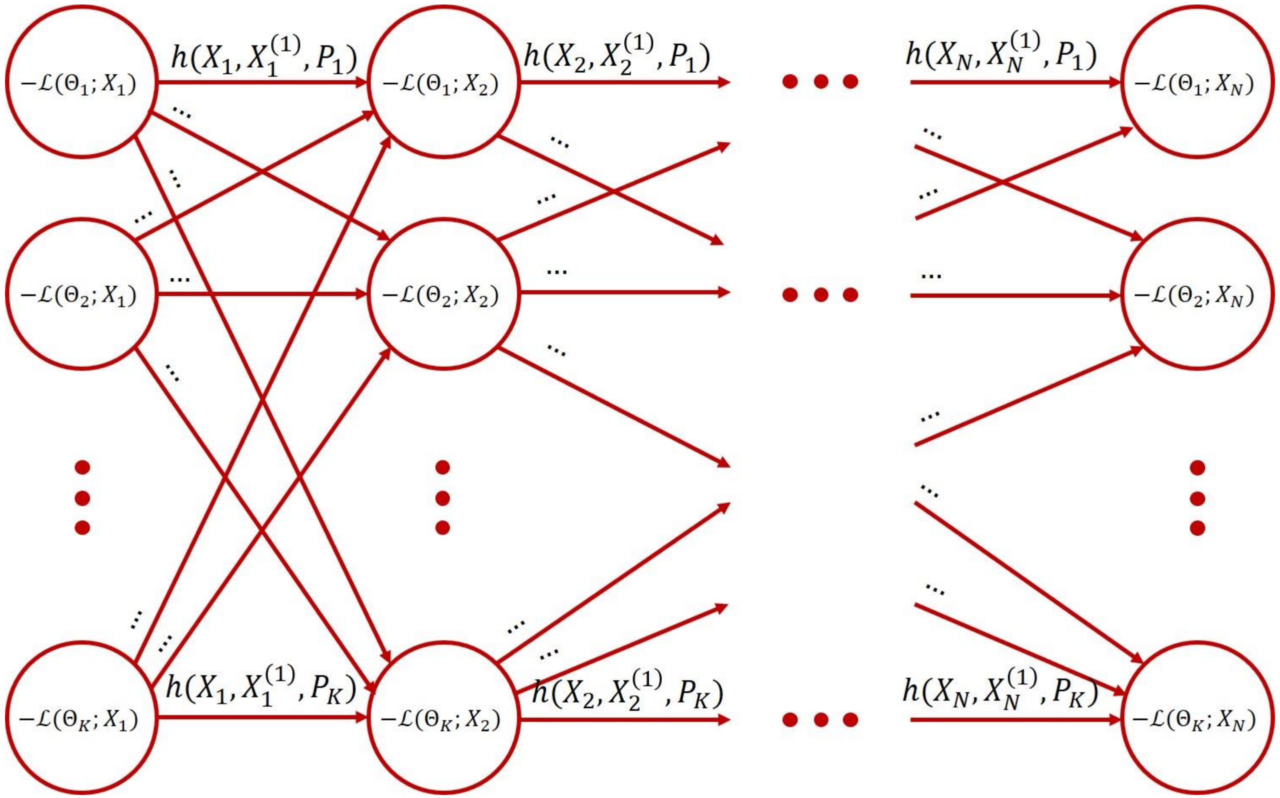
initialize cluster assignments $\boldsymbol{P}$ and cluster parameters $\boldsymbol{\Theta}$
while not stationarity
E-step: cluster assignments $\rightarrow \boldsymbol{P}$
M-step: parameter updates $\rightarrow \boldsymbol{\Theta}$
endwhile
return $\boldsymbol{P}, \boldsymbol{\Theta}$

E-M Style Approach

E-Step: Cluster Assignment

$$\min_{P} \sum_{k=1}^{K} \sum_{X_n \in P_k} \left(\overbrace{-\mathcal{L}(\Theta_k; X_n)}^{\text{log likelihood}} + \overbrace{\beta \mathbb{1}\{X_n^{(1)} \notin P_k\}}^{\text{spatial consistency}} \right)$$

Enumerate all possible assignments of the subregions to the clusters (infeasible!)



Strategy

Convert to a minimum cost path finding task from subregion 1 to N

- **Node: negative log likelihood of that point being assigned to a given cluster**
- **Edge: determined by the function whether the nth geographic object and its nearest neighbor belong to the same cluster**

M-Step: Cluster Parameter Updates

Toeplitz graphical lasso

$$\sum_{X_n \in P_k} \mathcal{L}(\Theta_k; X_n) = -|P_k|(\log\det\Theta_k + \mathrm{tr}(S_k\Theta_k)) + C,$$

$$\min_{\Theta_k \in \mathcal{T}} \quad -\log\det\Theta_k + \mathrm{tr}(S_k\Theta_k) + \frac{1}{|P_k|}\|\lambda \odot \Theta_k\|_{1,\mathrm{off}}$$

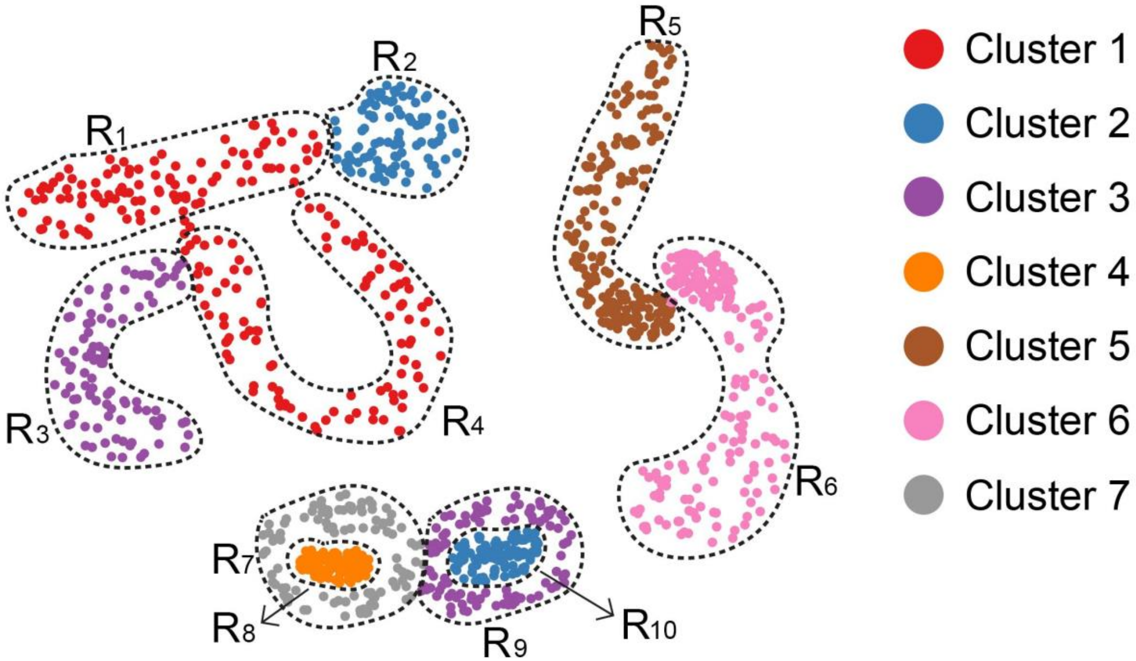
Strategy

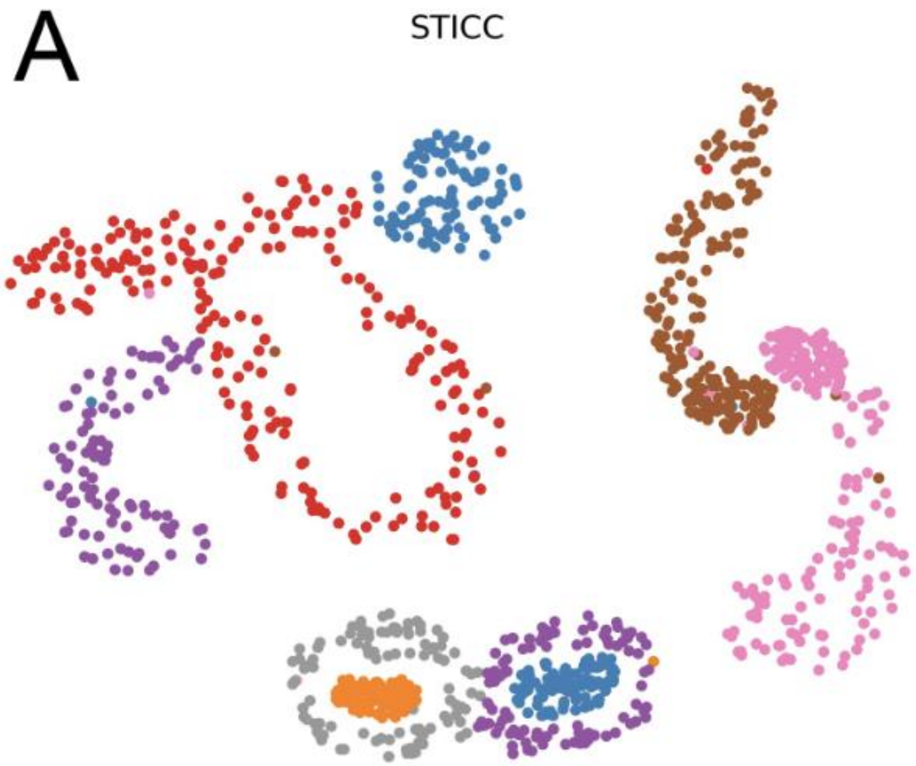
Solved using an alternating direction method of multipliers (ADMM) algorithm

Boyd, S., Parikh, N., Chu, E., Peleato, B. and Eckstein, J., 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1), pp.1-122.

Experiment I – A Synthetic Example

Ground Truth





Cluster	Attribute 1 μ	Attribute 1 θ	Attribute 2 μ	Attribute 2 θ	Attribute 3 μ	Attribute 3 θ	Attribute 4 μ	Attribute 4 θ	Attribute 5 μ	Attribute 5 θ
1	4	1	1	3	80	20	1000	350	999	3
2	5	1	7	3	30	20	900	350	992	3
3	6	1	2	3	20	20	600	350	1005	3
4	1	1	3	3	100	20	700	350	1003	3
5	3	1	6	3	60	20	800	350	999	3
6	7	1	4	3	70	20	400	350	998	3
7	2	1	5	3	40	20	500	350	1008	3

Experiment I – A Synthetic Example

	Cluster R	Cluster β	Adjusted rand index	Macro-F1	Join count ratio
STICC	1	3	0.894	0.954	0.851
	2	3	0.931	0.973	0.878
	3	3	**0.960**	**0.984**	0.901
	4	3	0.574	0.550	0.822
	3	0	0.947	0.977	0.888
	3	1	0.952	0.981	0.896
	3	5	0.818	0.771	0.886
Baseline clustering methods	*K*-Means		0.799	0.735	0.544
	CURE		0.0006	0.053	–
	Spatial K-Means		0.830	0.744	0.881
	GMM		0.703	0.850	0.685
	DBSCAN (*radius* = 1250, *minPts* = 25)		0.327	–	0.933
	Spatially constrained multivariate clustering		0.629	0.546	**0.936**

- Adjusted rand index and Macro-F1 can measure *accuracy* of clustering results
- Join count ratio can measure *spatial contiguity* of clustering results

Experiment II – Check-in Point Classification

New York Check-In Dataset

- Each POI contains at least 10 check-in points
- Only two fields: *week* and *hour* are taken into account when clustering

Objective

Cluster POIs into home/work, or home/work/gym categories

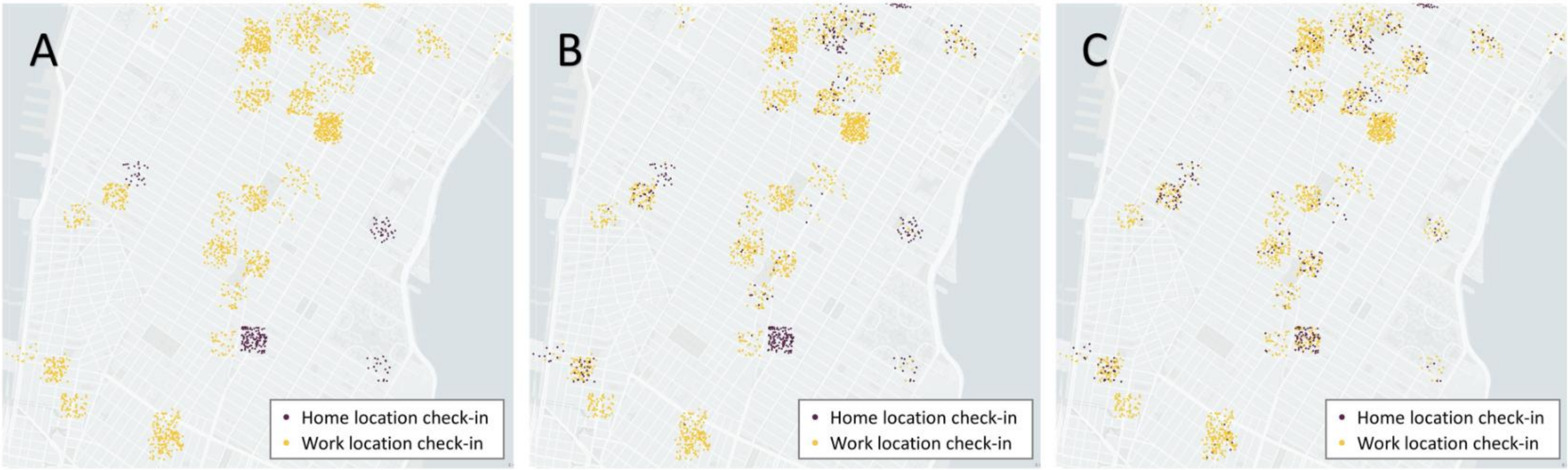
Experiment II – Check-in Point Classification

New York Check-In Dataset

- Each POI contains at least 10 check-in points
- Only two fields: *week* and *hour* are taken into account when clustering

Objective

Cluster POIs into home/work, or home/work/gym categories



Ground Truth STICC K-Means

Experiment II – Check-in Point Classification

Home/work classification

	Cluster				
	R	β	Adjusted rand index	Macro-F1	Join count ratio
STICC	1	3	0.390	0.806	0.829
	2	3	0.355	0.792	0.804
	3	3	0.433	0.823	0.834
	4	3	0.495	0.844	0.860
	4	0	0.445	0.823	0.822
	4	1	0.464	0.834	0.841
	4	5	**0.514**	**0.850**	**0.871**
Traditional clustering	*K*-Means		0.085	0.321	0.493
	CURE		0.015	0.578	0.700
	Spatial-Kmeans		0.080	0.384	0.492
	GMM		0.023	0.587	0.690

Home/work/gym classification

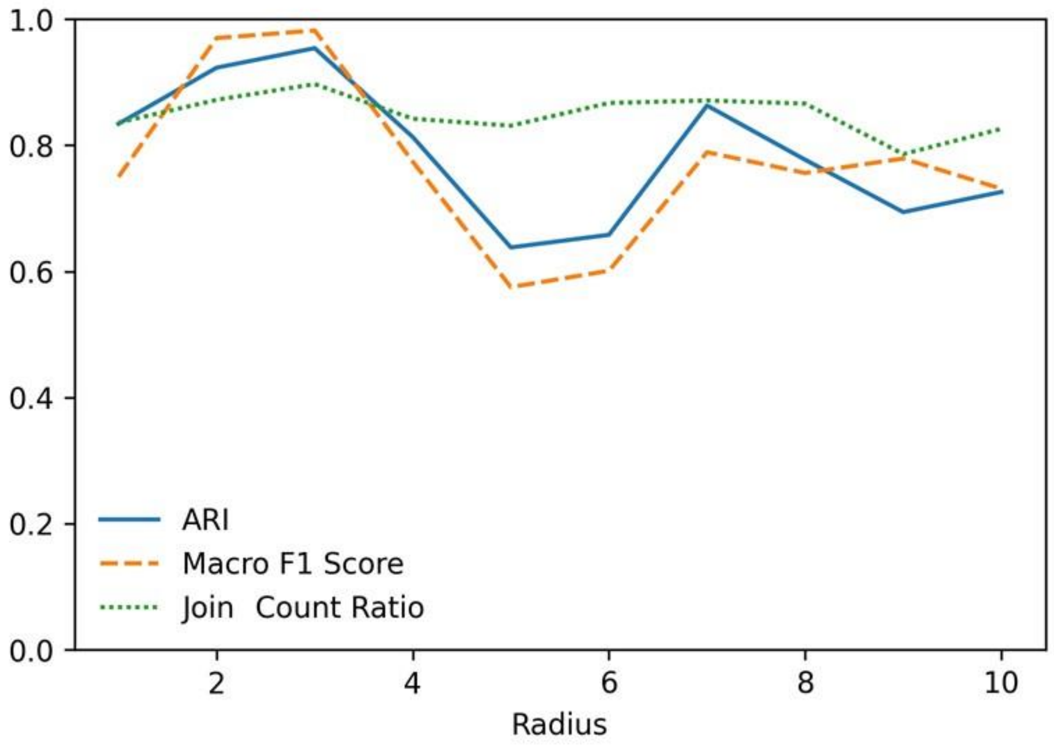
	Cluster				
	R	β	Adjusted rand index	Macro-F1	join count ratio
STICC	1	3	0.289	0.476	0.700
	2	3	0.204	0.482	0.647
	3	3	0.269	0.500	0.672
	4	3	0.298	0.508	0.700
	4	0	0.251	0.495	0.641
	4	1	0.273	0.502	0.669
	4	5	**0.335**	**0.510**	**0.712**
Traditional clustering	*K*-Means		0.041	0.352	0.675
	CURE		0.077	0.294	0.597
	Spatial-Kmeans		0.080	0.397	0.670
	GMM		0.065	0.416	0.603

Influences of Parameters

Four input parameters: K, β, λ, R

K – number of clusters

R – number of nearest neighbors in subregions



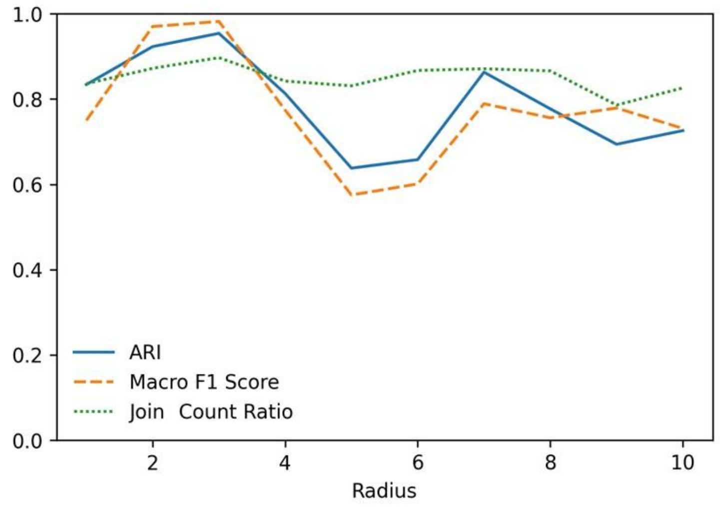
Influences of Parameters

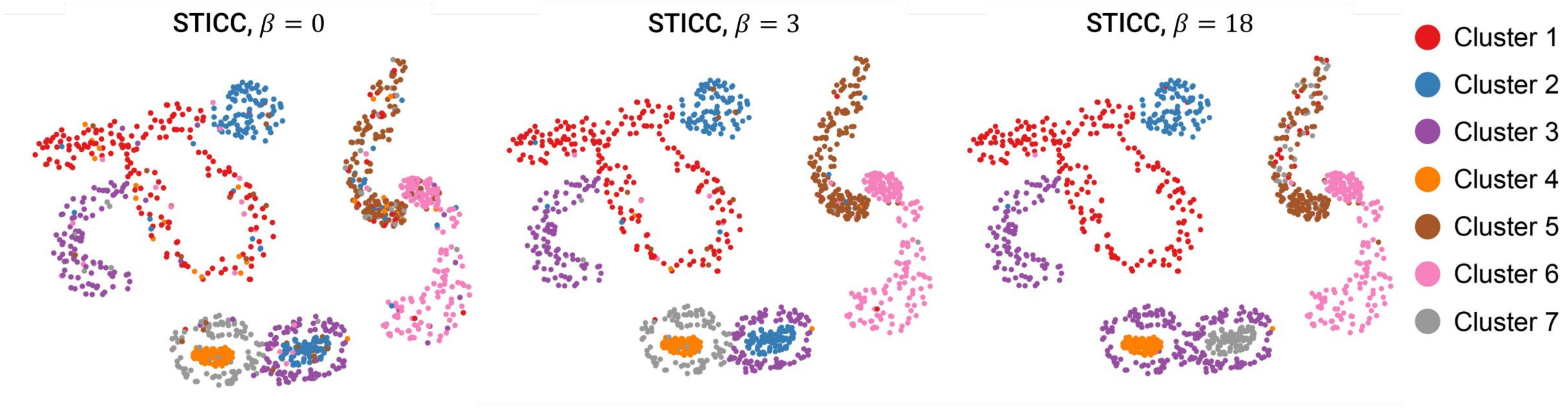
Four input parameters: K, β, λ, R

K – number of clusters

R – number of nearest neighbors in subregions

β – penalty costs of the spatial consistency





λ– level of sparsity

Clustering Result Interpretation

Network analysis approaches can be used for evaluating the properties of each cluster

	Attribute A	Attribute B	Attribute C	Attribute D	Attribute E
Cluster 1	0.00	0.00	0.50	0.83	0.00
Cluster 2	0.00	0.00	0.50	0.83	0.25
Cluster 3	0.83	0.00	0.08	0.83	0.58
Cluster 4	0.42	0.00	0.25	0.58	0.92
Cluster 5	0.17	1.00	0.00	0.42	0.67
Cluster 6	0.83	0.42	0.17	0.58	0.00
Cluster 7	0.00	0.17	0.92	0.33	0.83

Betweenness centrality of attributes in different clusters of the synthetic dataset

Clustering Result Interpretation

Network analysis approaches can be used for evaluating the properties of each cluster

	Attribute A	Attribute B	Attribute C	Attribute D	Attribute E
Cluster 1	0.00	0.00	0.50	0.83	0.00
Cluster 2	0.00	0.00	0.50	0.83	0.25
Cluster 3	0.83	0.00	0.08	0.83	0.58
Cluster 4	0.42	0.00	0.25	0.58	0.92
Cluster 5	0.17	1.00	0.00	0.42	0.67
Cluster 6	0.83	0.42	0.17	0.58	0.00
Cluster 7	0.00	0.17	0.92	0.33	0.83

Betweenness centrality of attributes in different clusters of the synthetic dataset

Attribute A is important in determining cluster 3, 4, and 6

Contributions

1. A novel spatial clustering method that considers both spatial and aspatial features for multivariate repeated geographic pattern discovery (RGPD)

2. The reliability and effectiveness of the proposed method is validated through synthetic experiments and real-world applications

3. The join count statistics is used to measure the spatial dependence of the clustering result

Code and Dataset

The data and codes that support the findings of this study are available on the Github repository:

https://github.com/GeoDS/STICC

README.md

License BSD 2-Clause

Geospatial Data Science Lab
UW-Madison

STICC: A multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity

GeoDS Lab, Department of Geography, University of Wisconsin-Madison.

Table of Contents

Citation

If you use this algorithm in your research or applications, please cite this source:

Kang, Y., Wu, K., Gao, S., Ng, I., Rao, J., Ye, S., Zhang, F. and Fei, T. STICC: A multivariate spatial clustering method for repeated geographic pattern discovery with consideration of spatial contiguity. *International Journal of Geographical*

Thank You!

The end

Yuhao Kang, Ph.D. student

Department of Geography

University of Wisconsin-Madison

yuhao.kang@wisc.edu www.kkyyhh96.site