IMPROVED NEAREST NEIGHBOR DENSITY-BASED CLUSTERING TECHNIQUES WITH APPLICATION TO HYPERSPECTRAL IMAGES

Claude Cariou[a], Kacem Chehdi[a] and Steven Le Moan[b]

a - Univ Rennes / Enssat - SHINE/TSI2M team
Institute of Electronics and Telecommunications of Rennes - CNRS UMR 6164, Lannion, France
claude.cariou@univ-rennes1.fr
b - Massey University, Center for Research in Image & Signal Processing, Palmerston North, New Zealand

Outline

- Introduction
- Proposed improvements:
  - structure of the nearest neighbor (NN) graph;
  - pointwise density model;
  - applicability to existing clustering methods.
- Experiments on hyperspectral images
- Conclusion

Introduction

- Clustering is a difficult problem in general:
  - Ill-posed: many clustering solutions exist;
  - Often requires hyperparameters;
  - The size of clustering problems is continuously increasing;
  - The dimensionality of data sets too;

- Some clustering methods can avoid specifying the number of clusters:
  - DBSCAN, OPTICS;
  - Mean Shift, Blurring Mean Shift;
  - Affinity Propagation
  - Convex clustering
  - **Nearest-neighbor density-based (NN-DB)**

Introduction

- Nearest Neighbor – Density Based (NN-DB) methods:
  - Modeseek [Duin *et al.*, *LNCS* **7626**, 2012; PRTools]
  - kNNClust [Tran *et al.*, *Comput. Stat. & Data Anal.* **51**, 2006]
  - KNN-DPC [after Rodriguez & Laio, *Science* **344**, 2014]
  - GWENN [Cariou & Chehdi, *Proc. IEEE IGARSS*, 2016]

- NN-DB methods show interesting properties for clustering purposes:
  - deterministic
  - require just one parameter: the # of nearest neighbors
  - work well with non-convex clusters

Introduction

Notations:

- Dataset $\boldsymbol{\mathcal{X}} = \{\mathbf{x}_i\}_{i=\mathbf{1},\ldots,\boldsymbol{N}}$, $\mathbf{x}_i \in \mathbb{R}^n, N$: #objects
- Metric *d*: Euclidean distance, $d(\mathbf{x}_i, \mathbf{x}_j) = d_{ij} = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2$
- Number of NNs K, first assumed constant $\forall \mathbf{x}_i$
- Directed K NN graph: $\mathcal{G} = (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{N}}_K(\boldsymbol{\mathcal{X}}))$

	density estimate	references
Non-parametric model	$\left(\sum_{k \in \mathcal{N}_K(\mathbf{x}_i)} d_{ik}\right)^{-1}$	[Cariou & Chehdi, *Proc. IEEE IGARSS*, 2016]
	$\sum_{k \in \mathcal{N}_K(\mathbf{x}_i)} d_{ik}^{-1}$	[Cariou & Chehdi, SPIE RS Europe, 2018]
Parametric model	$\exp\left(-\frac{1}{K}\sum_{k \in \mathcal{N}_K(\mathbf{x}_i)} d_{ik}^2\right)$	[Du et al., *Knowl.-Based Systems*, 2016]
	$\sum_{k \in \mathcal{N}_K(\mathbf{x}_i)} \exp\left(-\frac{N.d_{ik}}{\sum_{i=1}^{N} d_{ij_i^K}}\right)$	[Geng et al., *Inform. Science*, 2018]
	$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \sum_{k \in \mathcal{N}_K(\mathbf{x}_i)} \exp\left(-\frac{d_{ik}^2}{2\sigma^2}\right)$	[Le Moan and Cariou, *Proc. IVCNZ*, 2018]

Introduction

Problem position:

Is there any better NN graph than the classical one?

Objectives:

- Improving NN-DB methods regarding:
  - The structure of the KNN graph;
  - The choice of the pointwise density model;
  - The generalization of these methods to variable-NN graphs.

- Focused methods [Cariou & Chehdi, *Proc. SPIE RS Europe,* 2017]
  - KNN-DPC
  - GWENN-WM

- Focused application: pixel clustering in hyperspectral images

Proposed improvements

1. **Variable-*K* NN graphs**

- Concept of *hubness* [Radovanovic et al., Mach. Learn. Res. 2010]
- Hubs are “popular” nearest neighbors among objects
- Hubs are closer than any other objects to their respective cluster center
- Hubs have inspired modifications of KNN graphs, i.e.

Mutual Nearest Neighbors (MNN) [Stevens et al., IEEE T-GRS 2017]:

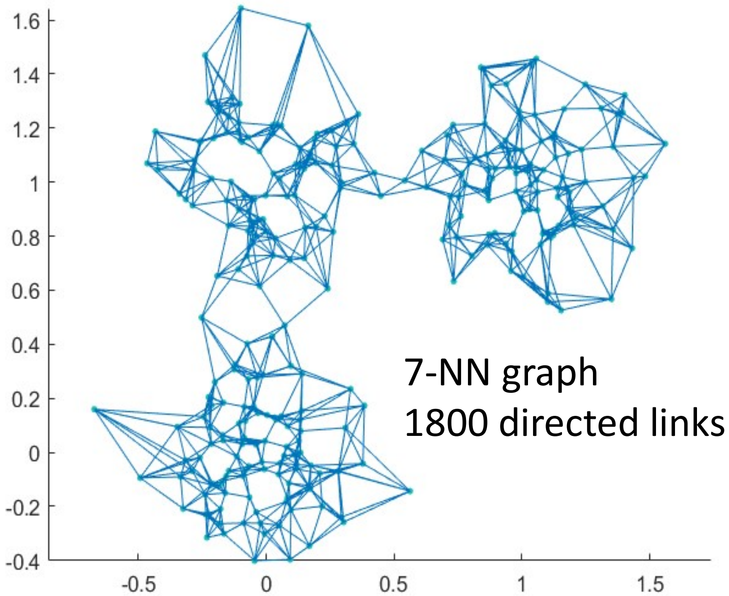
Remove edges to transform the original (directed) KNN graph into an undirected graph, such that

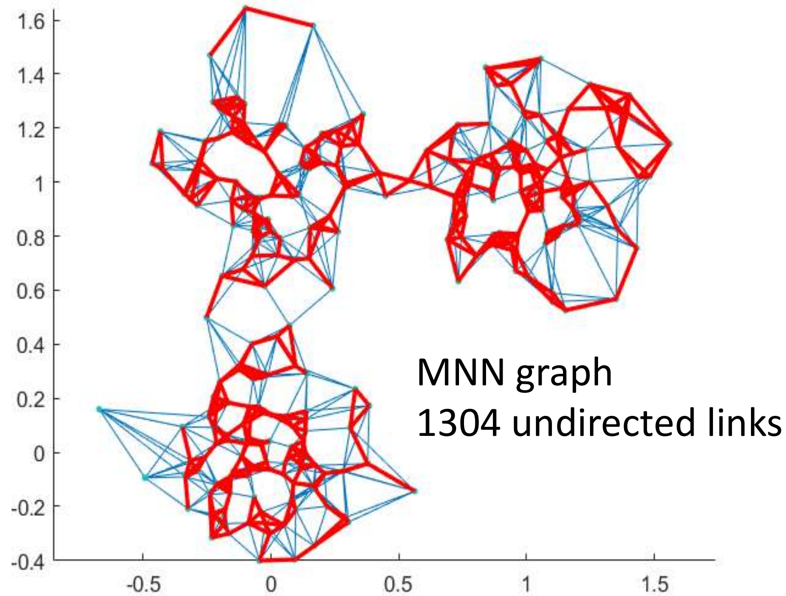
$(\mathbf{x}_i, \mathbf{x}_j)$ are connected iff $\left(\mathbf{x}_i \in \mathcal{N}_K(\mathbf{x}_j)\right) \wedge \left(\mathbf{x}_j \in \mathcal{N}_K(\mathbf{x}_i)\right)$

Proposed improvements

Consequences:

- The graph no longer has constant *K* outdegree
- Popular objects have larger outdegree than others





N = 300 2D-objects

Proposed improvements

2. Local density estimation

- Previous works based on constant *K* outdegree :

$$\rho(\mathbf{x}_i) = \frac{K}{\sum_{\mathbf{x}_j \in \mathcal{N}_K(\mathbf{x}_i)} d(\mathbf{x}_i, \mathbf{x}_j)} \qquad \rho(\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathcal{N}_K(\mathbf{x}_i)} \frac{\mathbf{1}}{d(\mathbf{x}_i, \mathbf{x}_j)}$$

- Proposed variable-*K* density model:

$$\rho(\mathbf{x}_i) = \frac{K_i}{d\left(\mathbf{x}_i, \mathcal{N}_{K_i}(\mathbf{x}_i)\right)} \;, \; 1 \leq i \leq N$$

Proposed improvements

3. Variable-K NN-DB clustering methods

- **KNN-Density Peak Clustering**
  - Find the unique neighbor of each object having the minimum distance among its neighbors of higher density;
  - Each object points to its nearest neighbor iteratively until convergence;
  - *No need for decision graph*.

- **GWENN-WM**
  - Rank the objects by decreasing local density;
  - Assign object's label as weighted mode label of *K*-nearest neighbors previously labelled;
  - If none of the *K* NNs of the current object is labeled yet, give it a new cluster label.

- Proposed improvement: replace regular KNN by MNN graph
  - → **MNN-DPC method**
  - → **GWENN-WM-MNN method**

Experimental study

- Application to hyperspectral images:
  - Large number of pixels ($N \sim 10^4$ to 10^7)
  - High dimensionality (dim ~ 10^2 to 10^3)

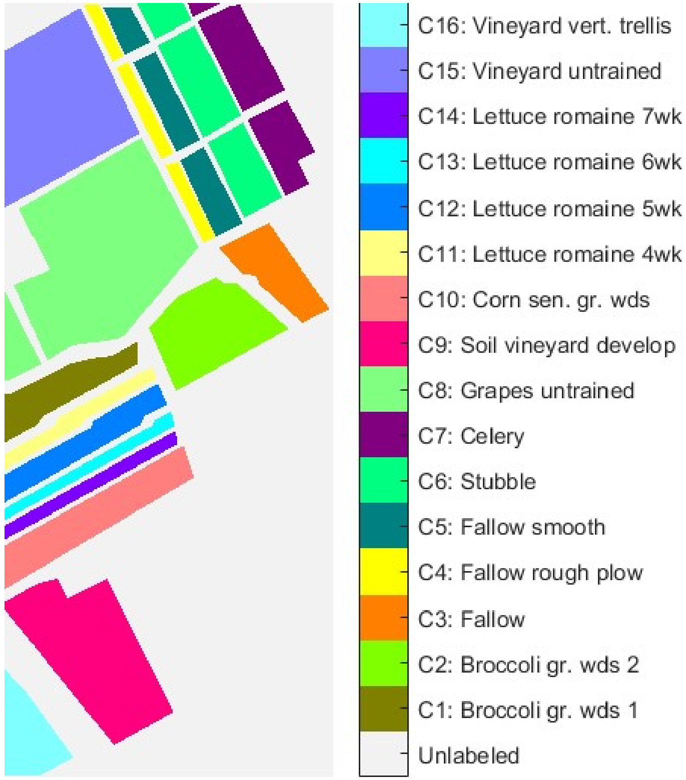
- A ground truth is available to assess the clustering results
  - Adjusted Rand Index [Hubert & Arabie, J. Classif. 1985]
  - Kappa index after confusion matrix reconditioning by Hungarian algorithm [Kuhn, 1955]

- Comparison with
  - KNN graph-based methods : KNN-DPC and GWENN-WM
  - Fuzzy C-Means [Bezdek, 1981]:
    - Two parameters: ***C***, *m* (here *m* = 2)
    - 20 random restarts
  - DBSCAN [Ester et al., Proc. KDD'96]:
    - Two parameters: *Eps*, *MinPts*

Experimental study

1. AVIRIS Salinas HSI dataset

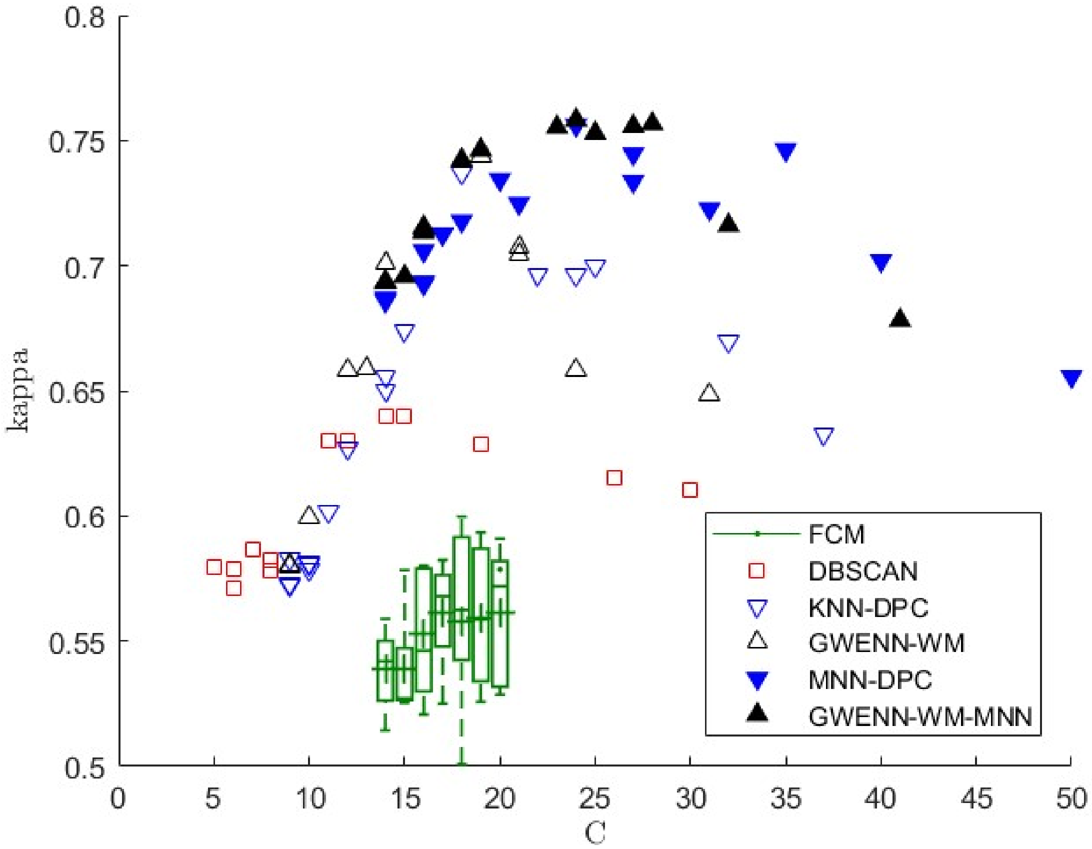
AVIRIS *Salinas* HSI
512x217 pixels
204 bands



Ground reference
54,129 pixels
16 classes

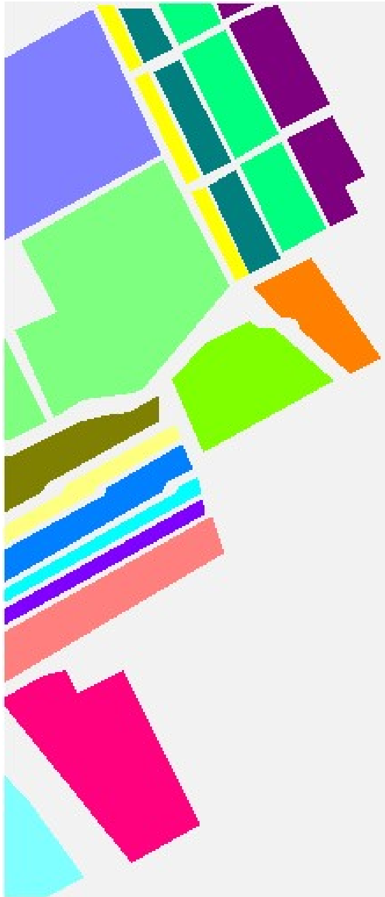
Experimental study

1. AVIRIS Salinas HSI dataset

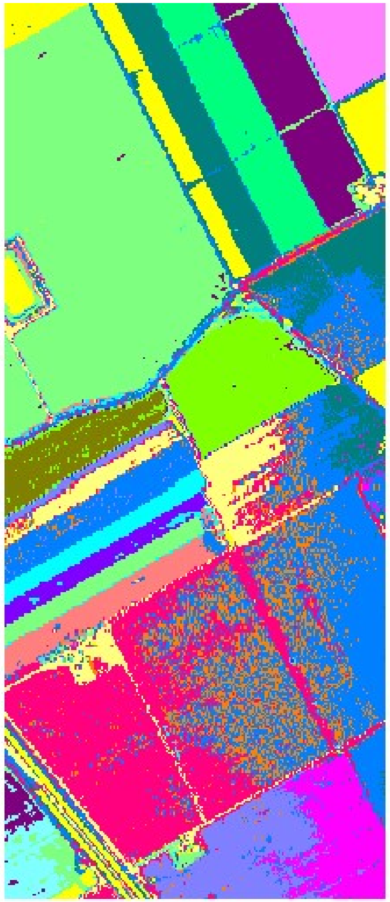


Experimental study

1. AVIRIS Salinas HSI dataset



Ground reference



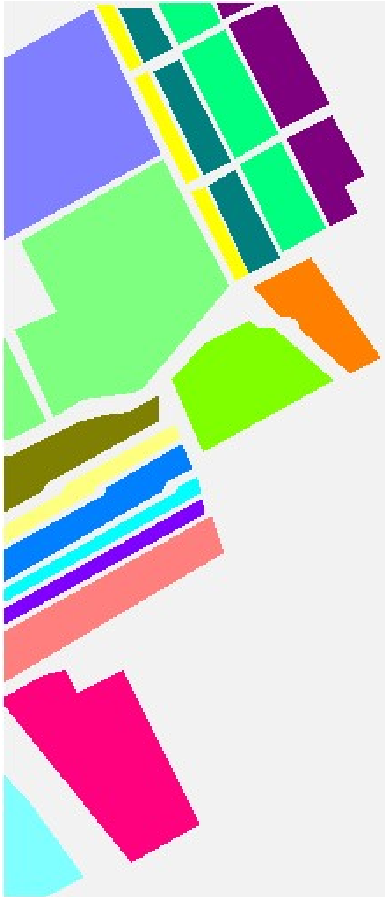
KNN-DPC
K=800, C=16 (18)
Kappa = 0.7373

MNN-DPC
K=1000, C=22 (24)
Kappa = 0.7561

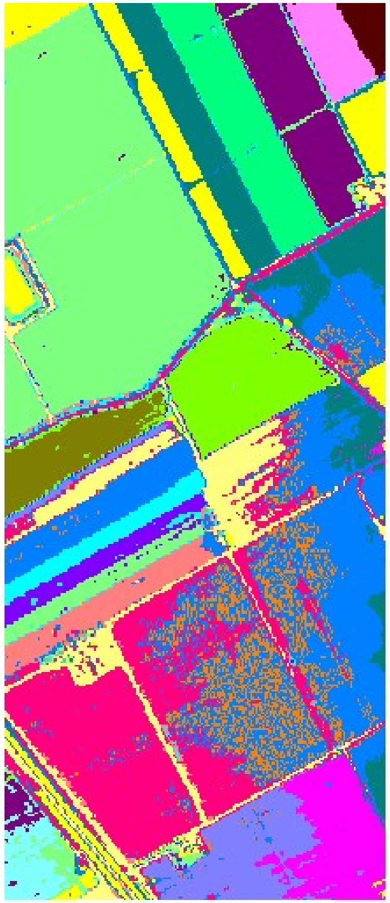
Salinas HSI

Experimental study

1. AVIRIS Salinas HSI dataset



Ground reference



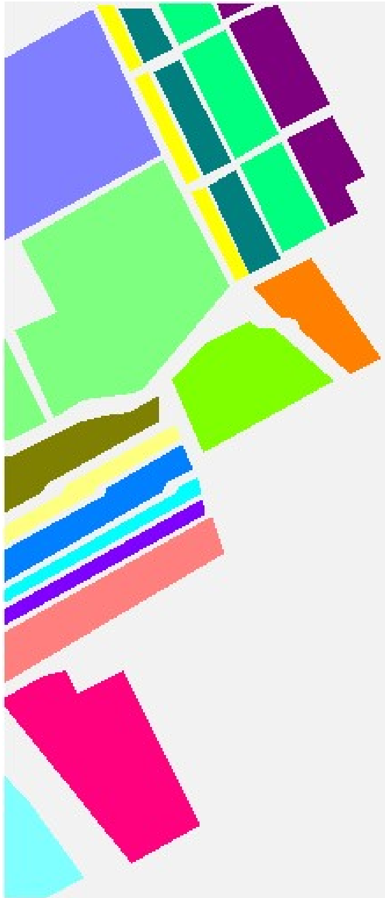
GWENN-WM-KNN
K=700, C=17 (19)
Kappa = 0.7441

GWENN-WM-MNN
K=900, C=22 (24)
Kappa = 0.7582

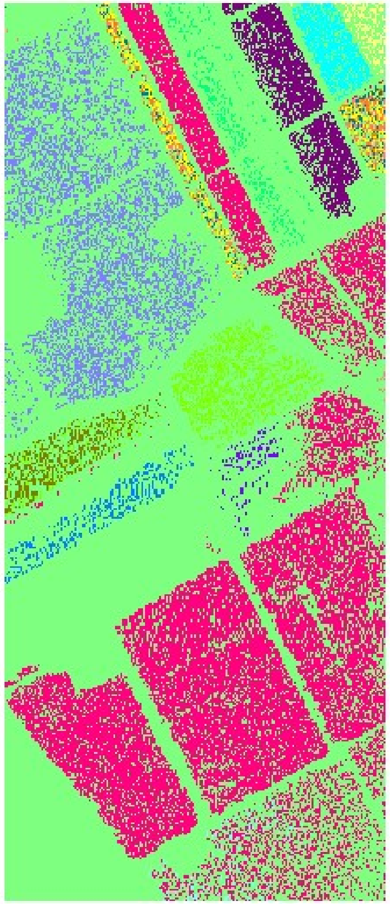
Salinas HSI

Experimental study

1. AVIRIS Salinas HSI dataset



Ground reference



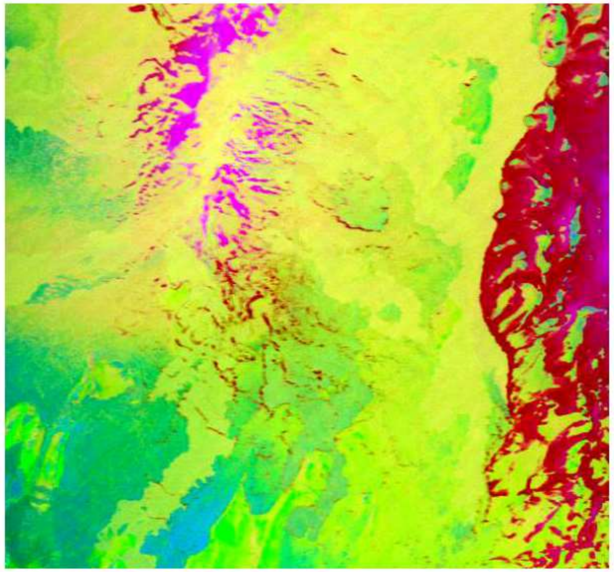
DBSCAN
Minpts=30, Eps=29
Kappa = 0.6399

FCM
C=24
Kappa = 0.5424

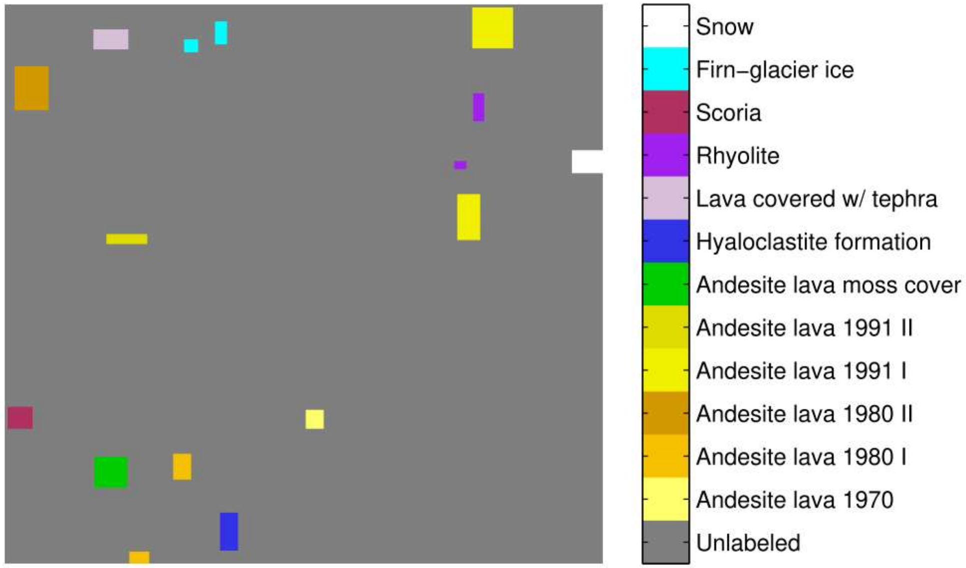
Salinas HSI

Experimental study

2. AVIRIS Hekla HSI dataset courtesy Prof. Jon Atli Benediktsson, U. Iceland



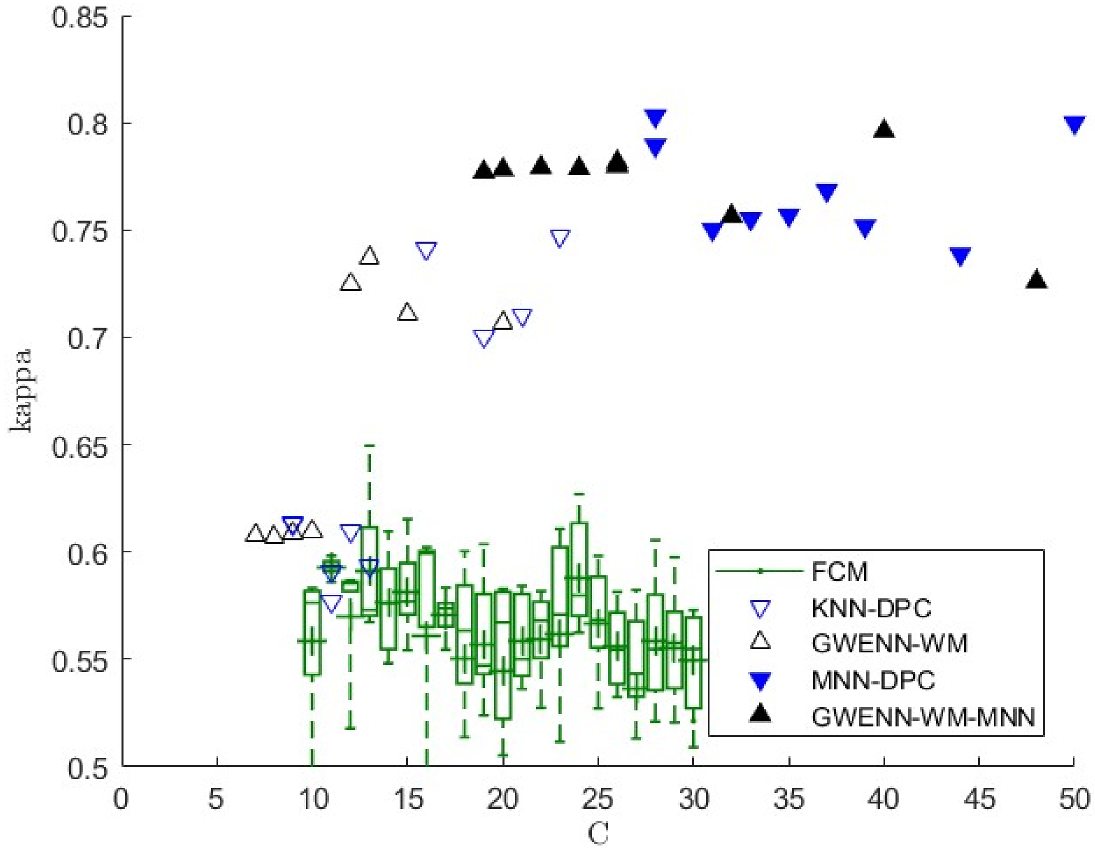
AVIRIS *Hekla* HSI
560x600 pixels
157 original bands
10 PC bands retained
after Minimum Noise Fraction



Ground reference
10227 pixels
12 classes

Experimental study

2. AVIRIS Hekla HSI dataset



Conclusion

- We have proposed a generalization of existing NN-DB clustering methods based on variable-*K* NN graphs;
- The NN graph was edge-pruned owing to the Mutual Nearest Neighbor (MNN) principle;
- MNN allows to highlight *hubs* in representation spaces, which are best suited for cluster unveiling than traditional medoids;
- NN-DB clustering methods are flexible enough to comply with variable-*K* NN graphs;
- Preliminary experiments on hyperspectral pixel clustering problems show the superiority of the proposed approach;
- Need for future investigation on modified NN graphs.