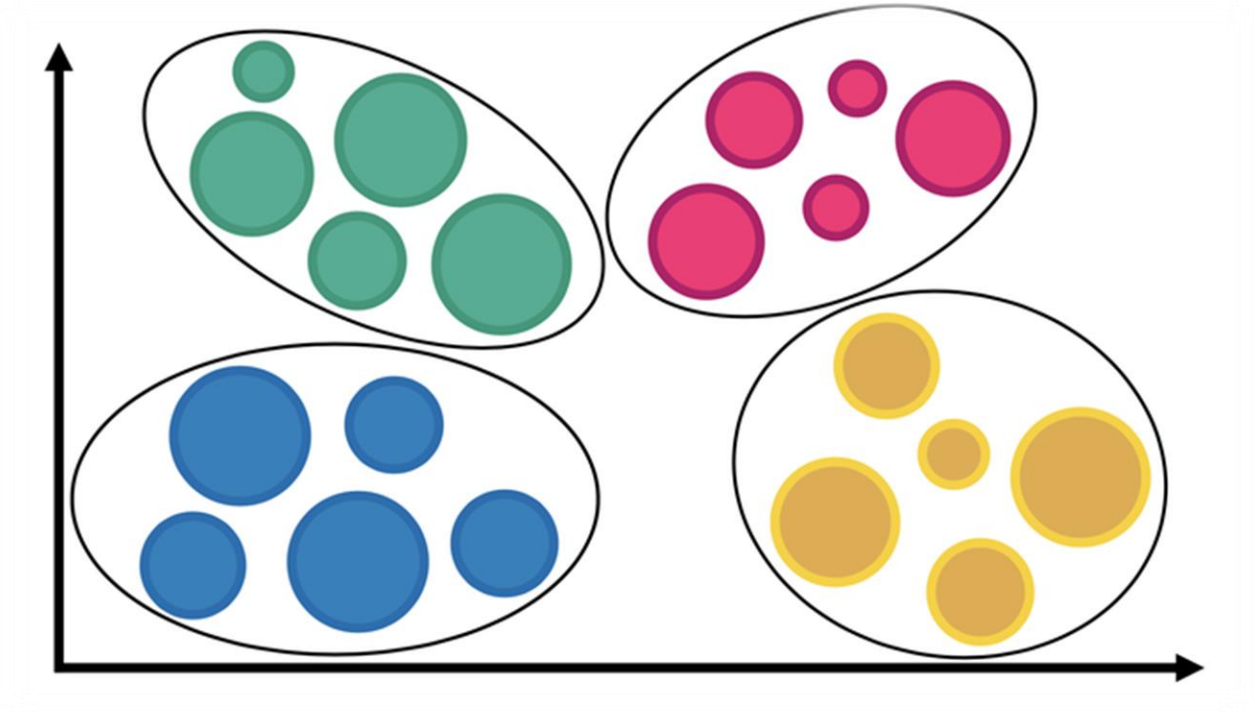

# CLUSTER ANANLYSIS IN DATA MINING

**Presented by : Naghmeh Pakgohar**

# What is clustering?

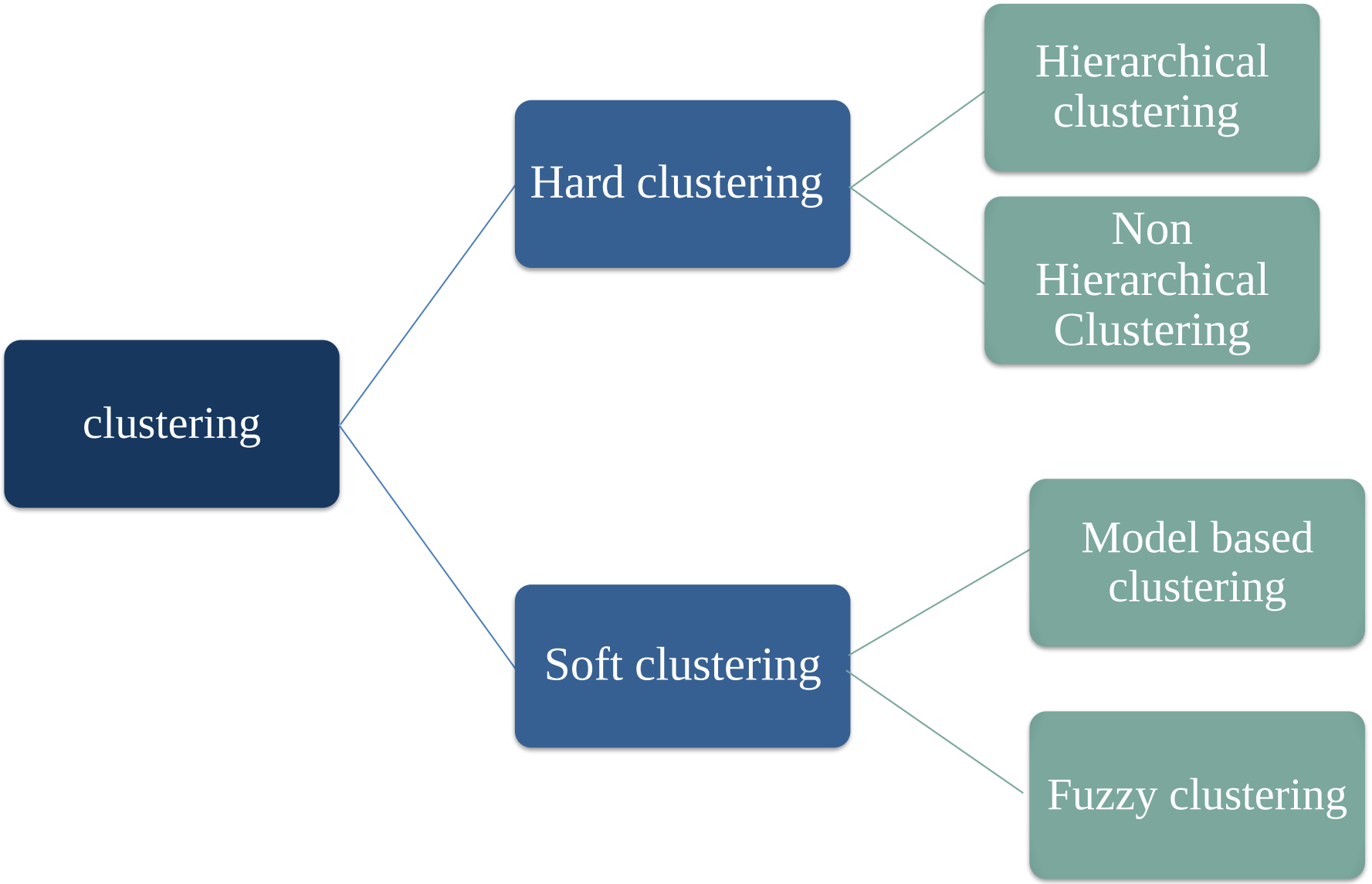


# What is clustering?

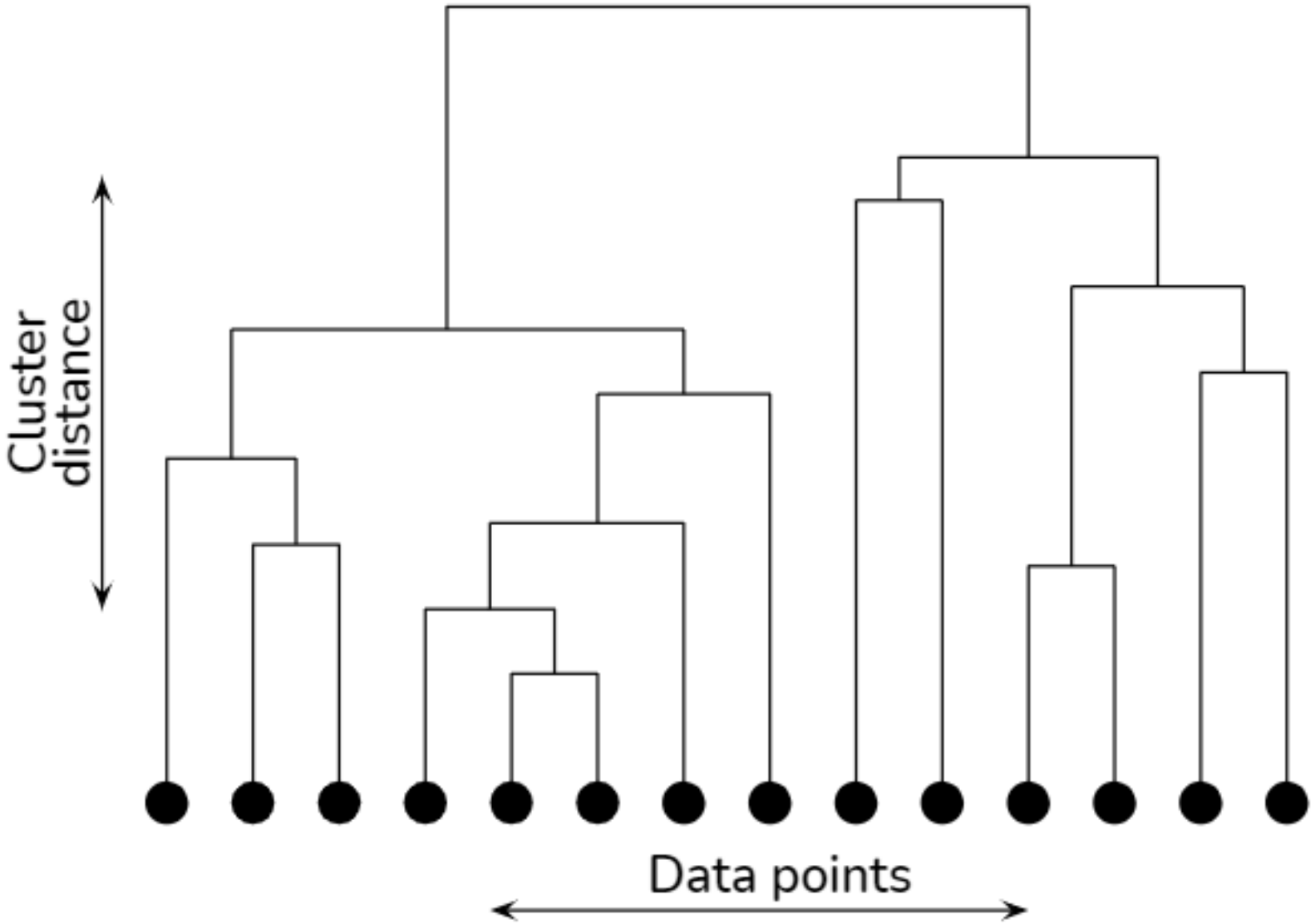
The aim of clustering

Summarize variation of the data

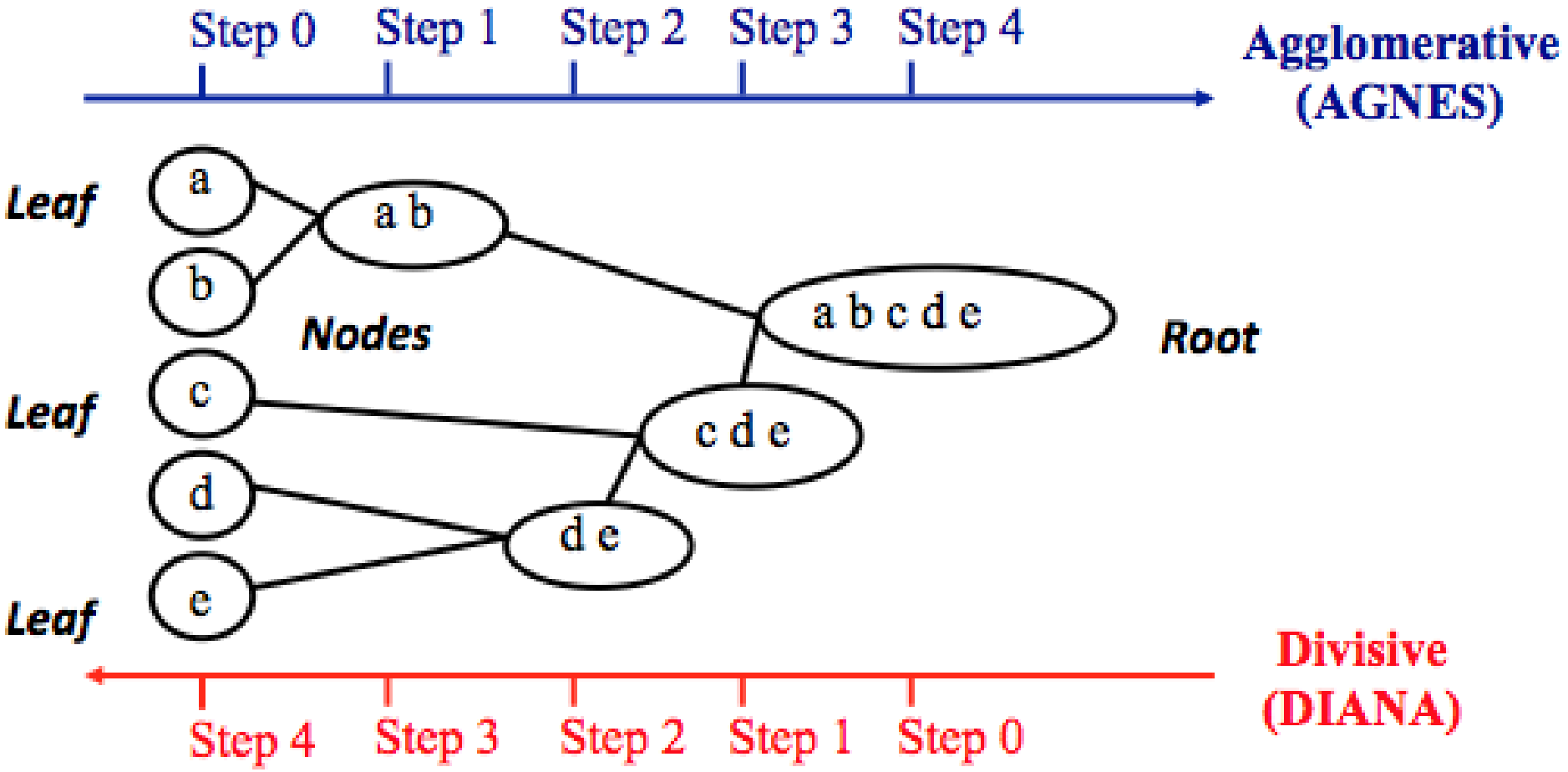
to get useful and practical information from multivariate data sets


clustering
Hard clustering
Hierarchical clustering
Non Hierarchical Clustering
Soft clustering
Model based clustering
Fuzzy clustering

# *Hierarchical clustering*

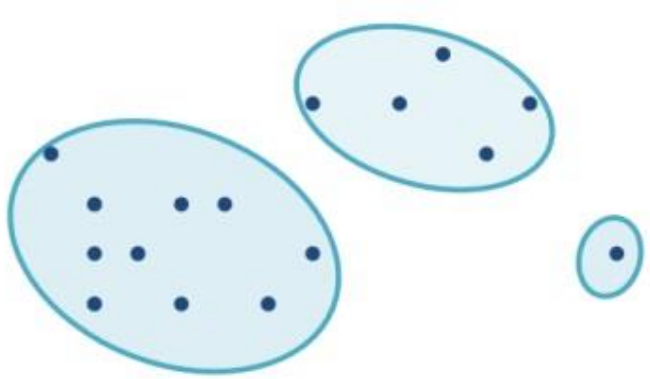


# *Hierarchical clustering*

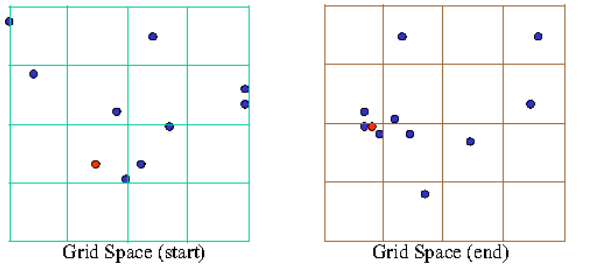


# Non-Hierarchical clustering

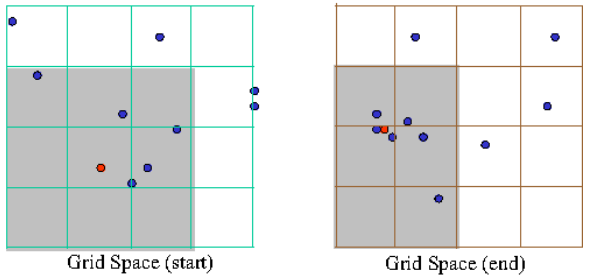
## Partitioning



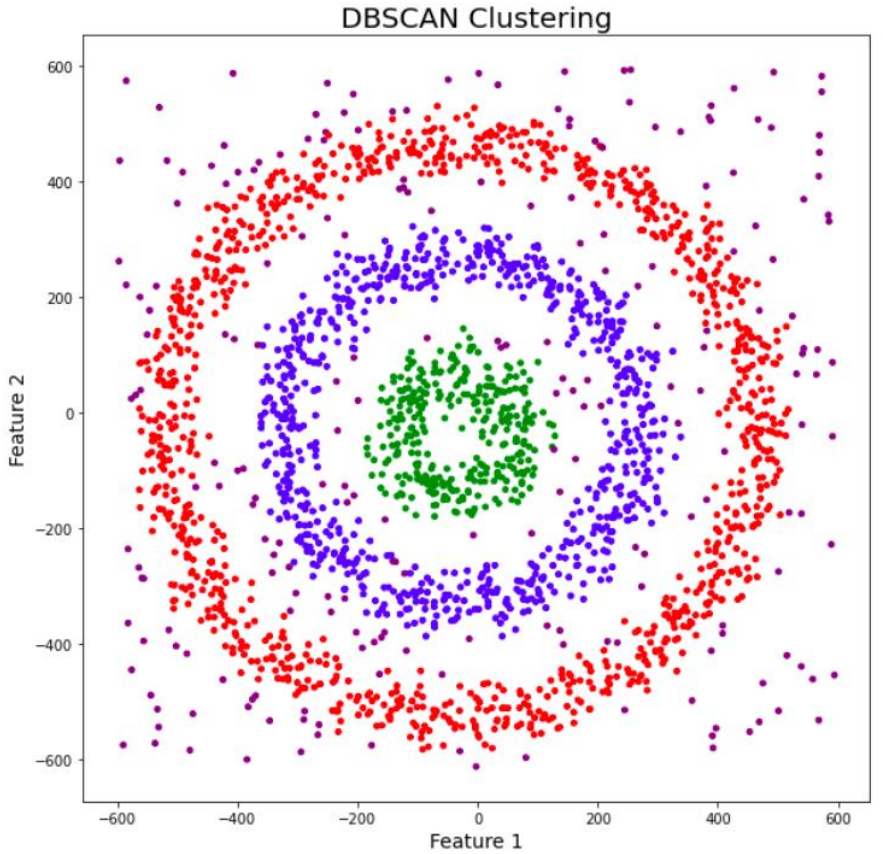
## Grid Based

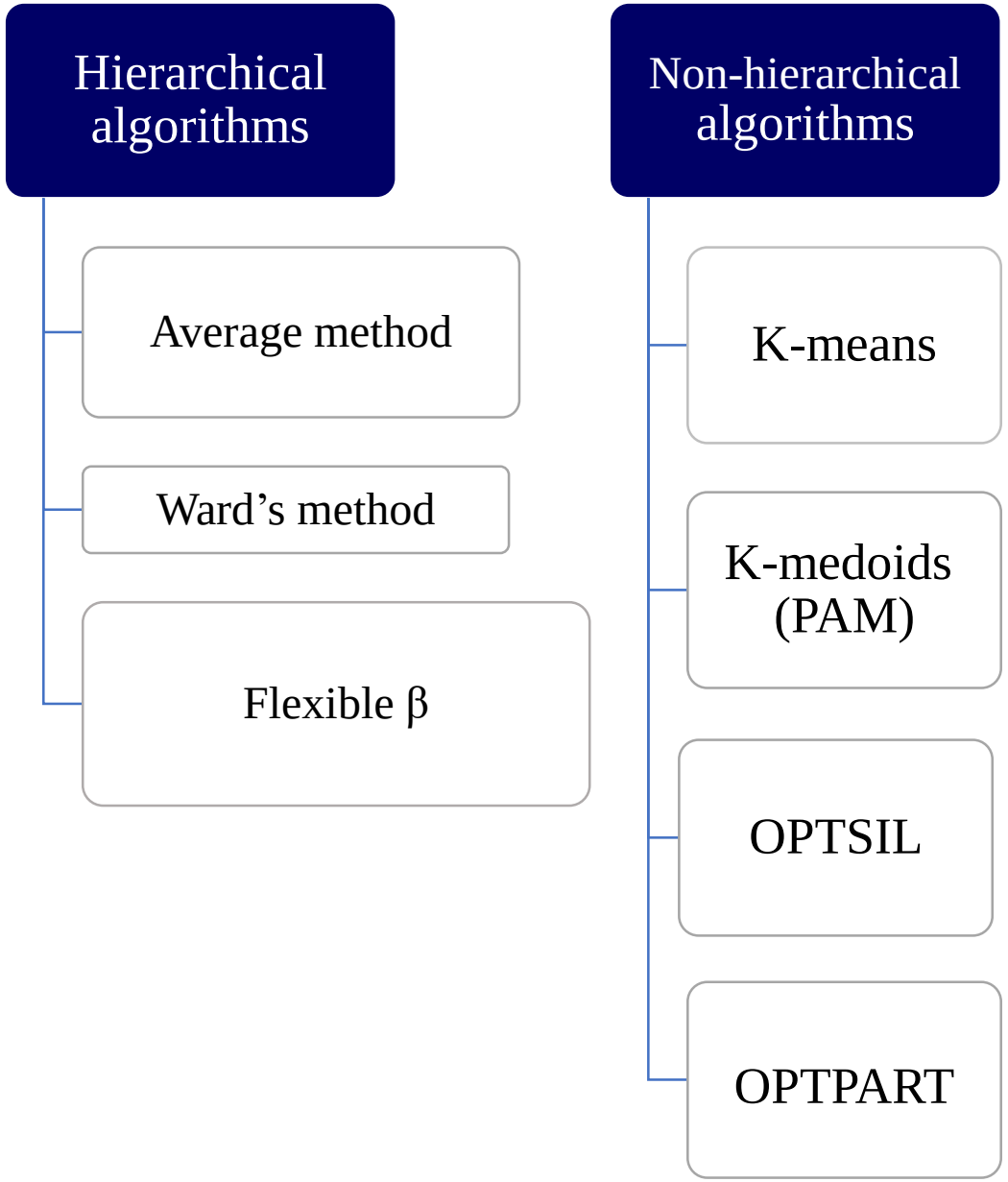



(A)

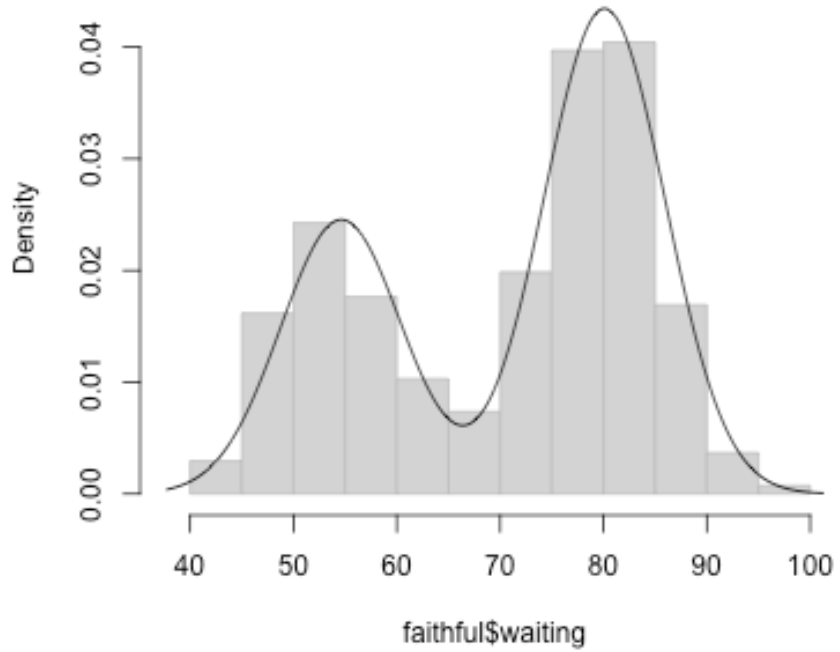


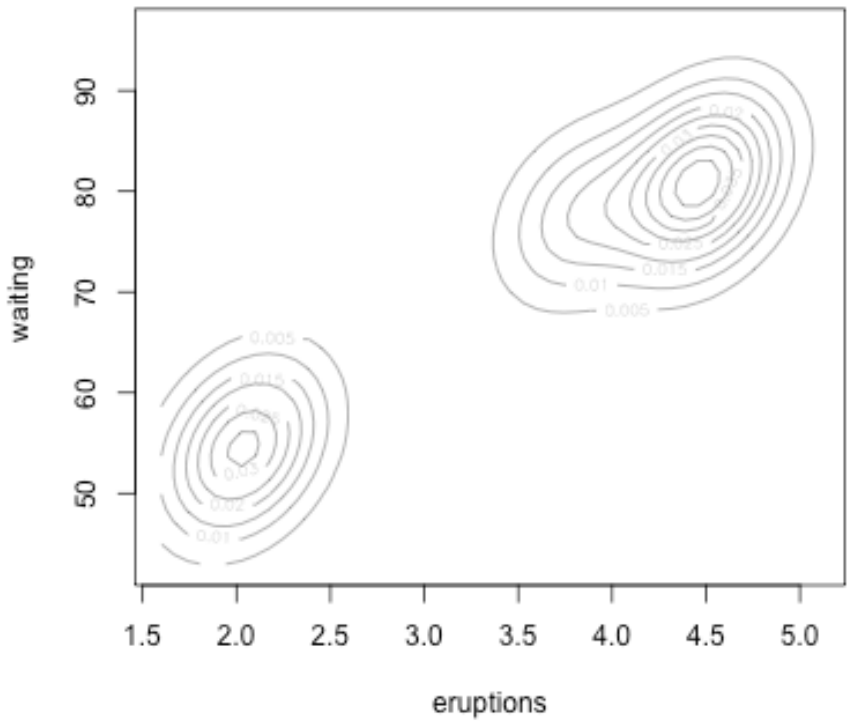

## Density Based



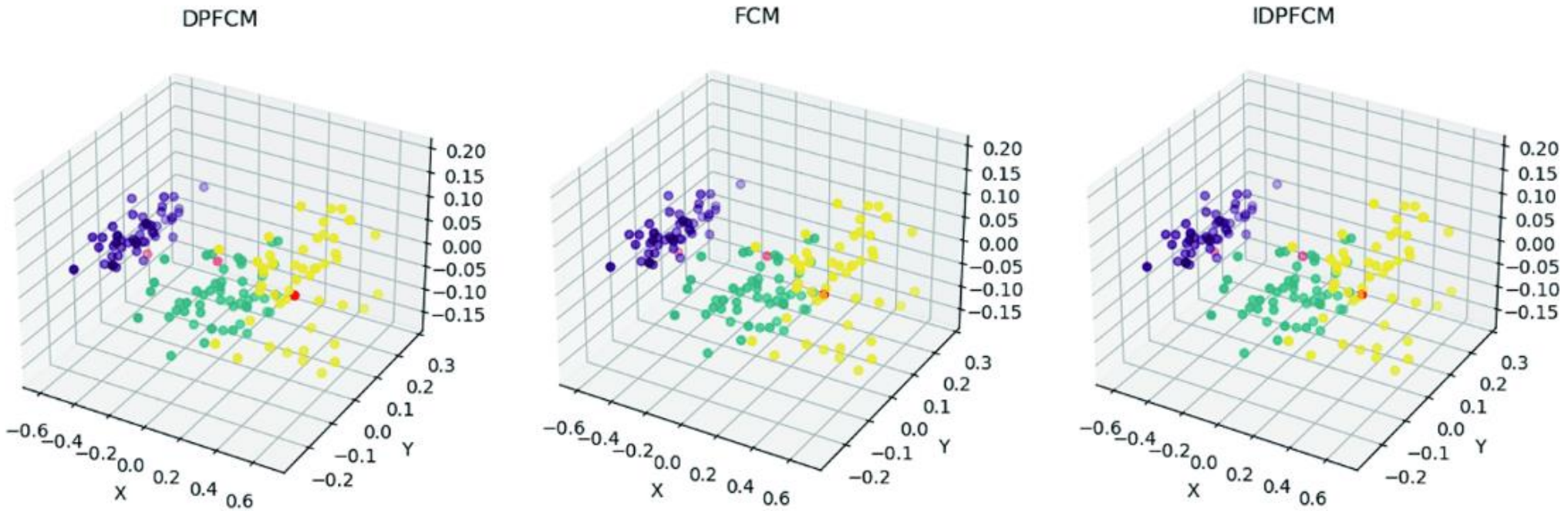

Hierarchical algorithms
Average method
Ward’s method
Flexible β
Non-hierarchical algorithms
K-means
K-medoids (PAM)
OPTSIL
OPTPART

# *Model based clustering*



# *Fuzzy clustering*




3D scatter map with privacy budget of 0.5

# *Main problems of clustering*

- 1- Lack of previous information about data
- 2- Different clustering methods are available
- 3. Objective decision

Even if the clustering algorithm is fixed, the grouping depends on the number of clusters.

Fundamental questions

Which  algorithm can achieve the best classification?

What is the optimal number of clusters?

# Goodness-of-clustering evaluators

## 1) Internal evaluators

- ✓ Internal evaluators use the characteristics of the clusters themselves to gauge effectiveness
  - Geometric evaluators → **e.g. Average Silhouette Width**
  - non-geometric evaluators → **e.g. Crispness**

## 2) External evaluators

- ✓ External evaluators compare the results of a classification with a previously established standard → **e.g. recovery of clusters embedded within simulated datasets**

Received: 10 February 2020 | Revised: 25 April 2021 | Accepted: 4 May 2021

DOI: 10.1111/jvs.13042

RESEARCH ARTICLE

Journal of Vegetation Science IAVS

# A comparative study of hard clustering algorithms for vegetation data

Naghmeh Pakgohar[1] | Javad Eshaghi Rad[1] | Gholamhossein Gholami[2] | Ahmad Alijanpour[1] | David W. Roberts[3]

[1]Department of Forestry, Faculty of Natural Resources, Urmia University, Urmia, Iran

[2]Department of Mathematics, Faculty of Science, Urmia University, Urmia, Iran

[3]Department of Ecology, Montana State University, Bozeman, MT, USA

**Correspondence**
Javad Eshaghi Rad, Department of Forestry,

## Abstract

**Questions:** Which clustering algorithms are most effective according to different cluster validity evaluators? Which distance or dissimilarity measure is most suitable for clustering algorithms?

**Location:** Hyrcanian forest, Iran (Asia), Virginia region forest, United States (North America), beech forests, Ukraine (Europe).

New Results

Follow this preprint

# Quantitative evaluation of internal clustering validation indices using binary datasets

Naghmeh Pakgohar, Attila Lengyel, Zoltán Botta-Dukát

**doi:** https://doi.org/10.1101/2023.08.09.552566







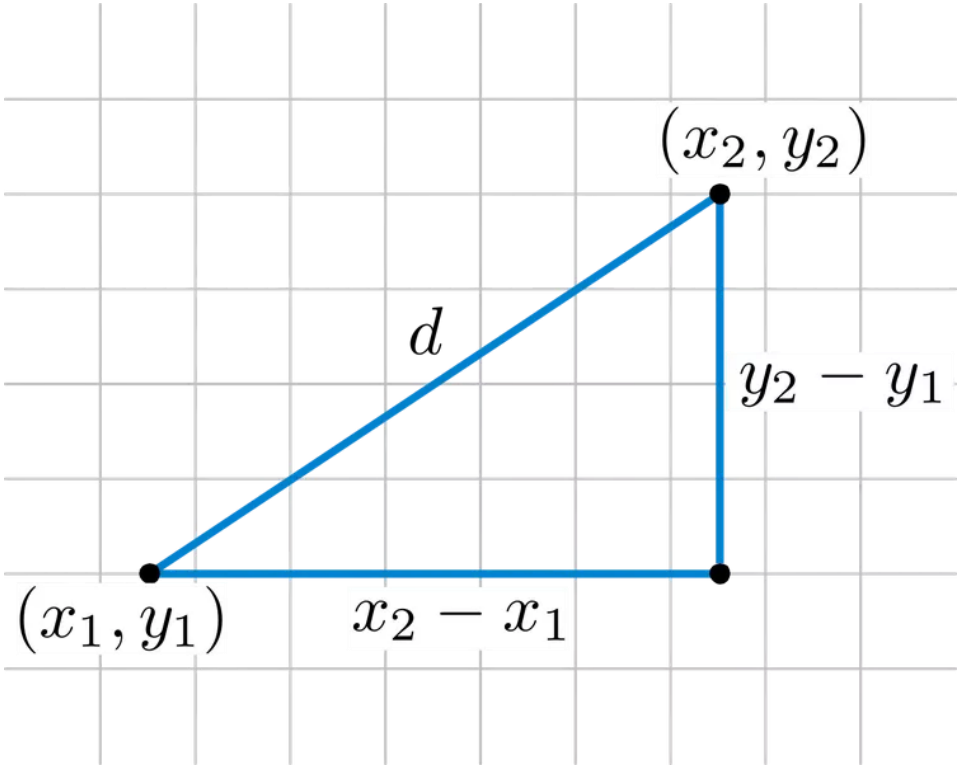
## Abstract

Different clustering methods often classify the same dataset differently. Selecting the 'best' clustering solution out of a multitude of alternatives is possible with cluster validation indices. The behavior of validity indices changes with the structure of the sample and the properties of the clustering algorithm. Unique properties of each index cause increasing or decreasing performance in some conditions. Due to the large variety of cluster validation indices, choosing the most suitable index concerning

***Hierarchical clustering***

***Cluster Analysis***

**Distance/dissimilarity indices**



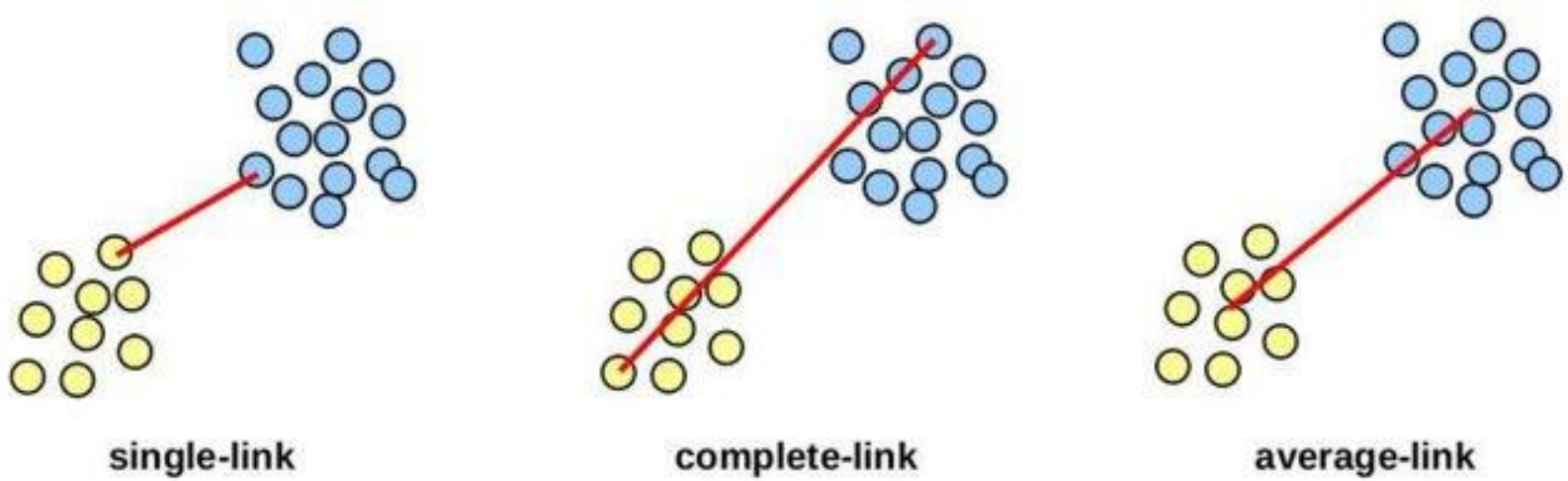

- ✓Sorensen Distance
- ✓Relative Sorensen Distance
- ✓Jaccard Distance
- ✓Euclidean Distance
- ✓Relative Euclidean Distance
- ✓ Squared Euclidean Distance
- ✓Correlation
- ✓Chi-squared Distance
- ✓Manhattan
- ✓Hellinger

# Distance/dissimilarity matrix

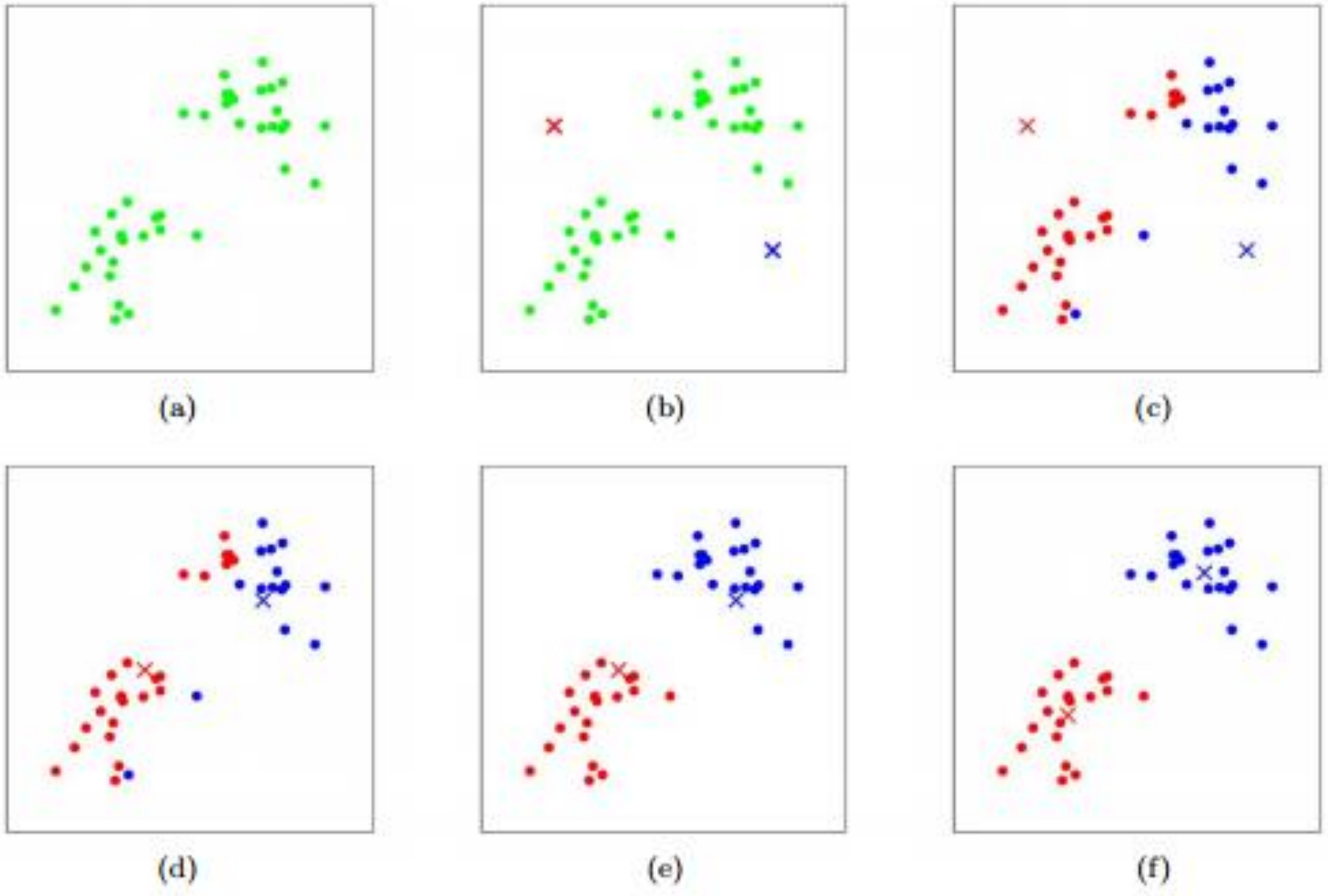
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 87.85 | 26.99 | 28.15 | 60.87 | 60.01 | 59.2 | 59.46 | 17.39 | 34.63 |
| 2 | 87.85 | 0 | 74.39 | 101.5 | 75.22 | 77.78 | 76.07 | 74.92 | 79.92 | 71.65 |
| 3 | 26.99 | 74.39 | 0 | 42.47 | 45.83 | 44.32 | 46.36 | 45.47 | 21.09 | 22.2 |
| 4 | 28.15 | 101.5 | 42.47 | 0 | 80.5 | 80.14 | 79.05 | 79.12 | 29.28 | 55.91 |
| 5 | 60.87 | 75.22 | 45.83 | 80.5 | 0 | 34.71 | 38.11 | 35.27 | 54.3 | 39.63 |
| 6 | 60.01 | 77.78 | 44.32 | 80.14 | 34.71 | 0 | 28.93 | 26.35 | 56.33 | 33.38 |
| 7 | 59.2 | 76.07 | 46.36 | 79.05 | 38.11 | 28.93 | 0 | 23.34 | 55.06 | 36.84 |
| 8 | 59.46 | 74.92 | 45.47 | 79.12 | 35.27 | 26.35 | 23.34 | 0 | 55.05 | 37.38 |
| 9 | 17.39 | 79.92 | 21.09 | 29.28 | 54.3 | 56.33 | 55.06 | 55.05 | 0 | 30.27 |
| 10 | 34.63 | 71.65 | 22.2 | 55.91 | 39.63 | 33.38 | 36.84 | 37.38 | 30.27 | 0 |

# *linkage method*

- ✓ Nearest neighbor or single linkage
- ✓ Farthest neighbor or complete linkage
- ✓ Between-group average linkage (UPGMA)
- ✓ Centroid
- ✓ Median
- ✓ Ward’s method
- ✓ Flexible Beta



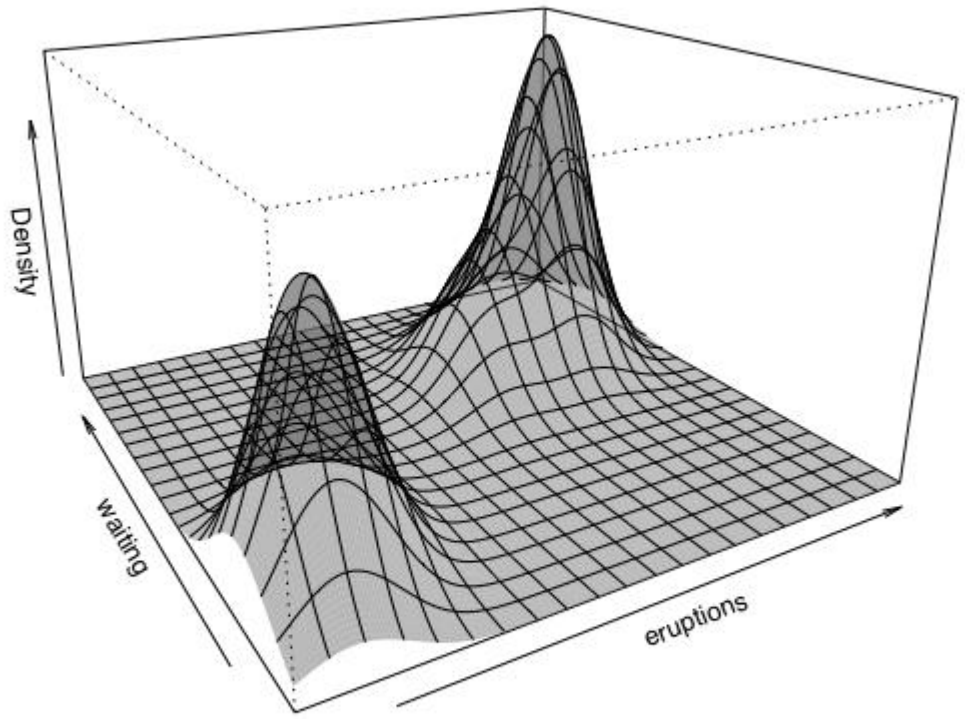
# *Kmeans*



Mclust

Finite Gaussian mixture modeling fitted

EM algorithm



# About me

Naghmeh Pakgohar
Postdoc fellow in Saskatchewan university,
Postdoc Position at Centre for Ecological Research,
Ph.D. of forestry(Quantitative ecology),
Interested in Vegetation Community, Spatial Pattern,
Environmental Ecosystem, Numerical Modeling


npakgohar@gmail.com

@Naghmeh.Pakgohar

https://scholar.google.com/citations?user=PkYZtCQAAAAJ&hl=en

Naghmeh Pakgohar