Project 4

Cluster analysis of Phishing URL:s

Florence Hugh, August Jonasson, Sofia Näslund

Unsupervised Learning - MT7050

2024-10-31

Introduction

- Phishing is a type of cyberattack where attackers trick individuals into providing sensitive information
- The dataset concerns URLs for phishing and non-phishing websites

Data cleaning

- Original data: 18 features and 2.5 million observations
- Removed features: "URL", "source", "who is data" and "domain age days"
- Sampling of the dataset: 1 500 samples
- Result: 14 features and 1 500 observations

Features

Binary Features	Explanation
"starts_with_ip"	Indicates if the URL starts with an IP address
"has_punycode"	Indicates if the URL contains punycode
"domain_has_digits"	Indicates if the domain contains digits
"has_internal_links"	Indicates if the URL contains internal links
"label"	Indicates whether the link is legitimate or phishing
Integer Features	**Explanation**
"url_length"	Number of characters in the URL
"dot_count"	Number of dots (‘.‘) in the URL
"at_count"	Number of at:s (‘@‘) in the URL
"dash_count"	Number of dashes (‘-‘) in the URL
"tld_count"	Number of top-level domains in the URL
"subdomain_count"	Number of subdomains in the URL
Continuous Features	**Explanation**
"url_entropy"	Randomness of the URL characters
"nan_char_entropy"	Randomness of non-alphanumeric characters in the URL
"digitletter_ratio"	Ratio of digits to letters in the URL

Dissimilarity measure

- Mixed data - categorical and numerical - how to deal with it?

Dissimilarity measure

- Mixed data - categorical and numerical - how to deal with it?

Gower's distance

Defined as

$$d_{ij} = \frac{\sum_k^D w_{ijk} d_{ijk}}{\sum_k^D w_{ijk}},$$

where w_{ijk} are the weights and d_{ijk} is the k:th feature dissimilarity between observations i and j.

For categorical features: $\mathrm{d}_{ijk} = \begin{cases} 1 & \text{if} \quad y_{ik} = y_{jk}, \\ 0 & \text{if} \quad y_{ik} \neq y_{jk}. \end{cases}$

For numerical features:

$$d_{ijk} = \frac{|y_{ik} - y_{jk}|}{R_k},$$

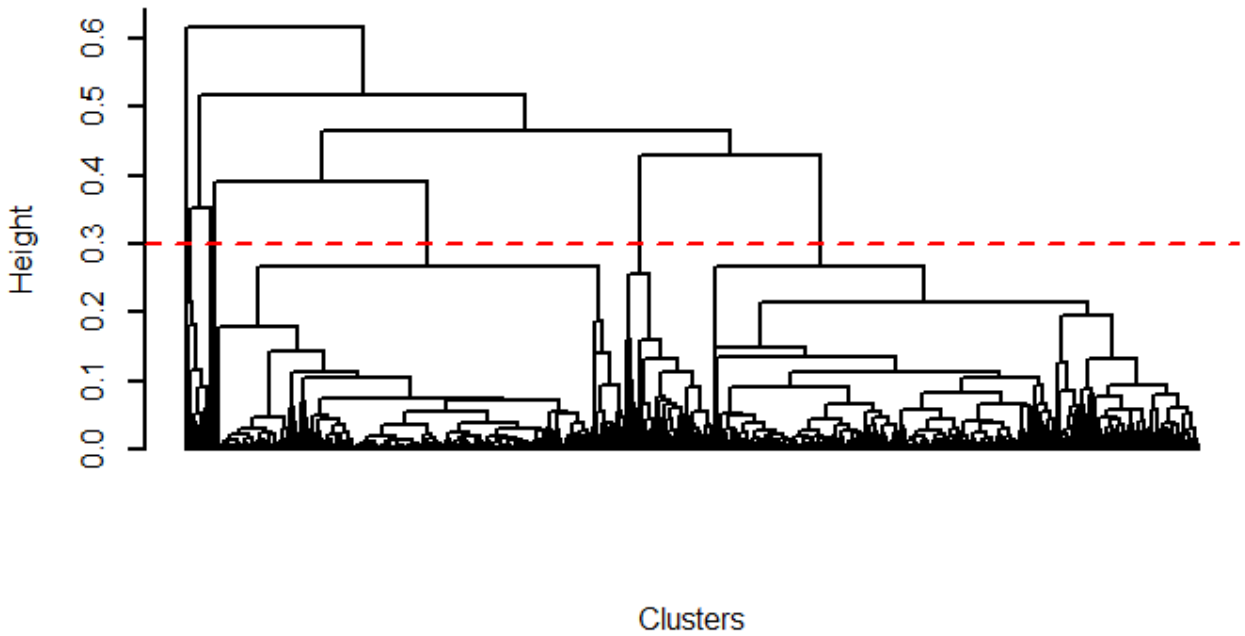
where R_k is the range of k:th feature, thus ensuring that $d_{ijk} \in [0, 1]$.

Exploratory Analysis

- Explore if there exists any natural cluster structure

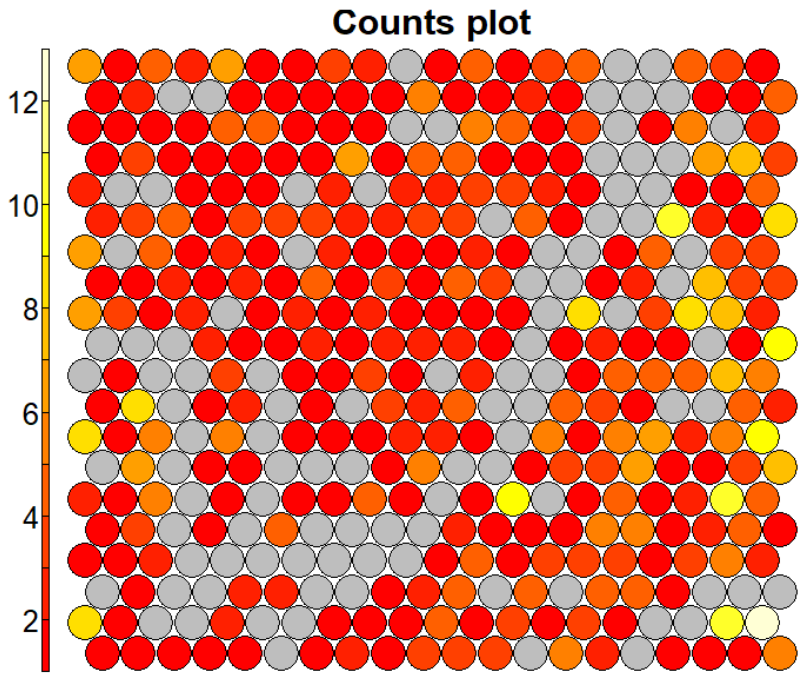
Exploratory Analysis

- Explore if there exists any natural cluster structure
- Hierarchical clustering - Complete linkage



Exploratory Analysis

- Self-organizing map



Dimensionality reduction

- Remove noise and "useless" features
- Preservation of distances and/or topology

Dimensionality reduction

- Remove noise and "useless" features
- Preservation of distances and/or topology

Non-metric MDS (nmMDS)

Minimize the stress function

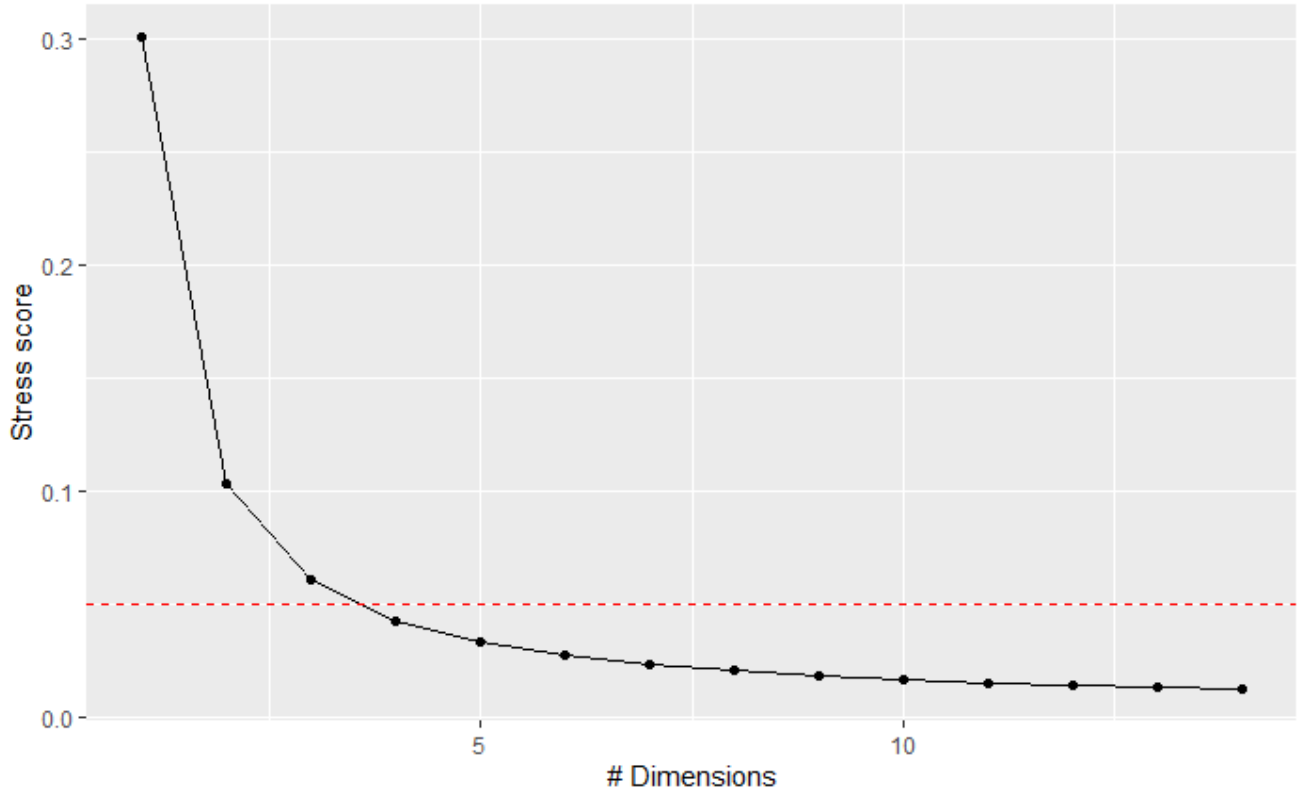
$$E_{nmMDS} = \sqrt{\frac{\sum_{i,j}^{N} w_{ij} |f(\delta(i,j)) - d_x(i,j)|^2}{c}},$$

where

- w_{ij} weights
- $\delta(i,j)$ rank order of dissimilarity between y_i and y_j
- $f(\cdot)$ monotonic (step-wise) regression against original dissimilarities
- d_x Euclidean distances in embedded space
- c is some constant that prevents collapse of solution

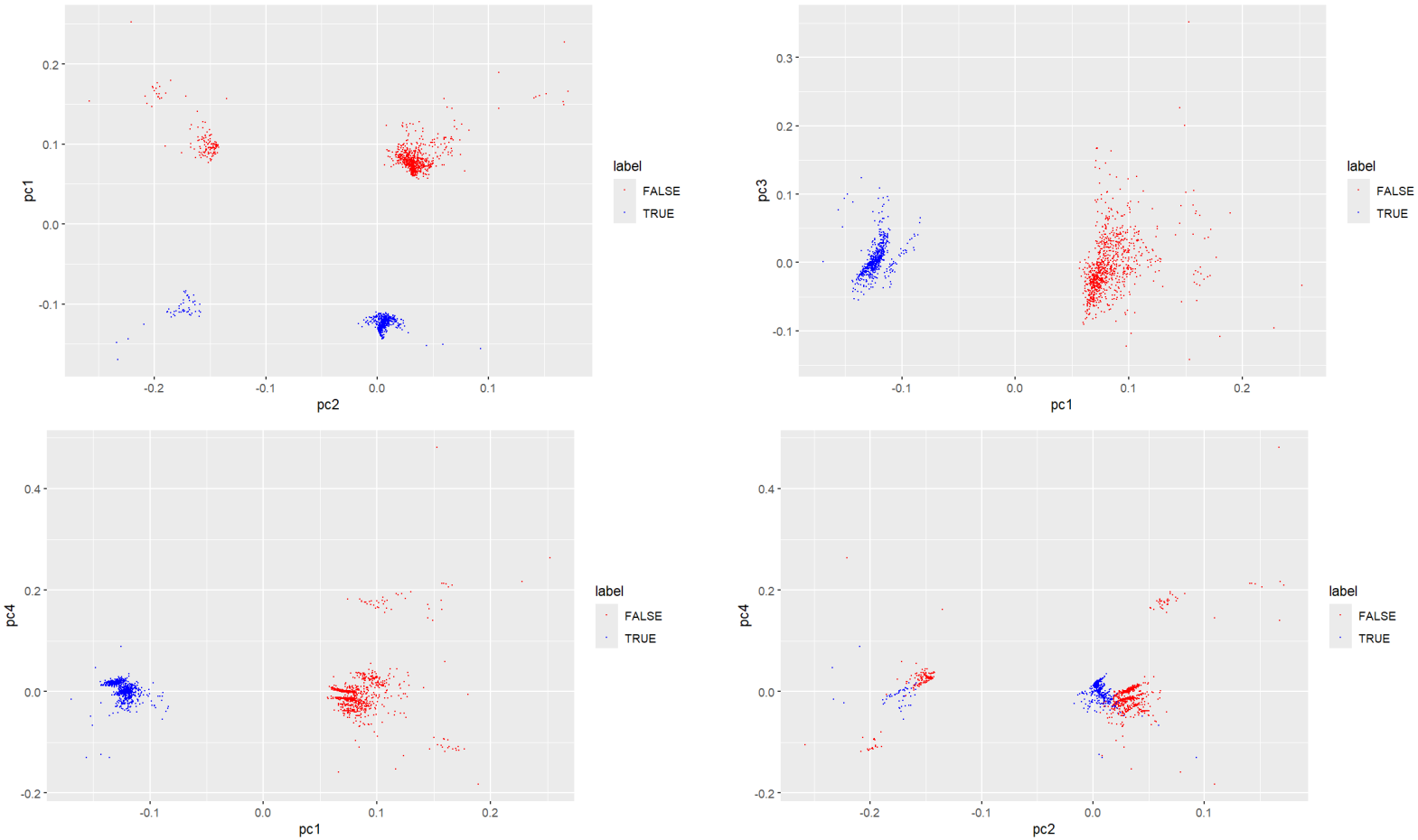
Dimensionality reduction

- Choose number of dimensions such that stress falls below 5 % (this is not validation)



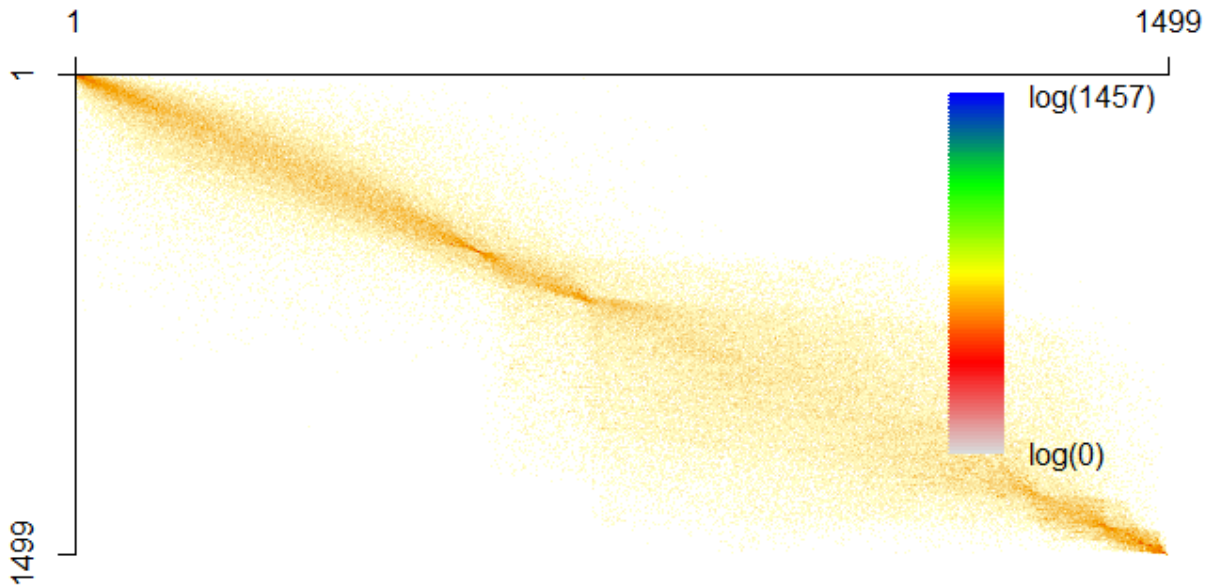
Dimension reduction

- Non-metric MDS pairs of principal coordinates



Validation of dimensionality reduction

- Co-ranking matrix
  - Sparse
  - Very mild intrusions/extrusion in local neighborhood
  - Less mild but still mild for higher order ranks
  - No mixing between near and far away neighbors (good for preserving clusters)

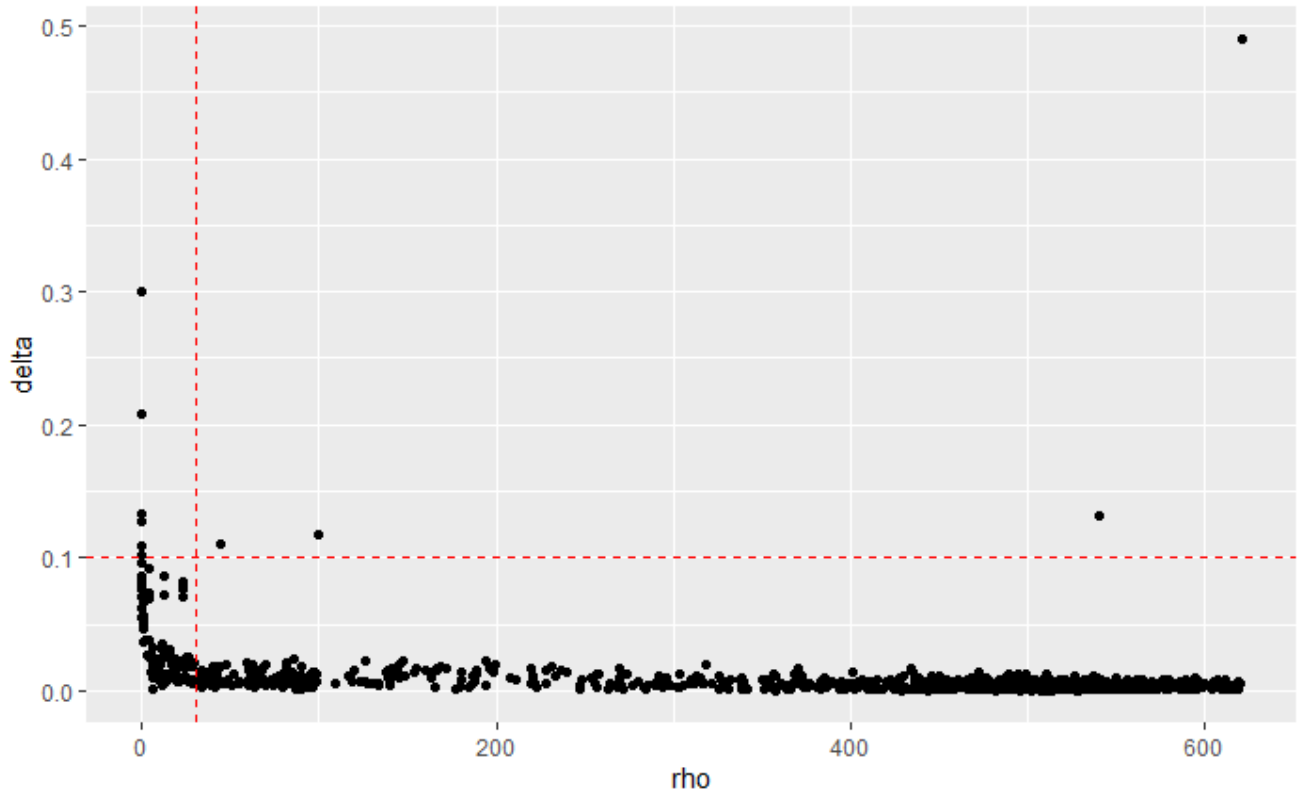


Clustering

- Clustering method: Density Peak Clustering (DPC)
  - Data points looks well separated, vary in density and non-linear shapes
  - Deal with noise and outliers

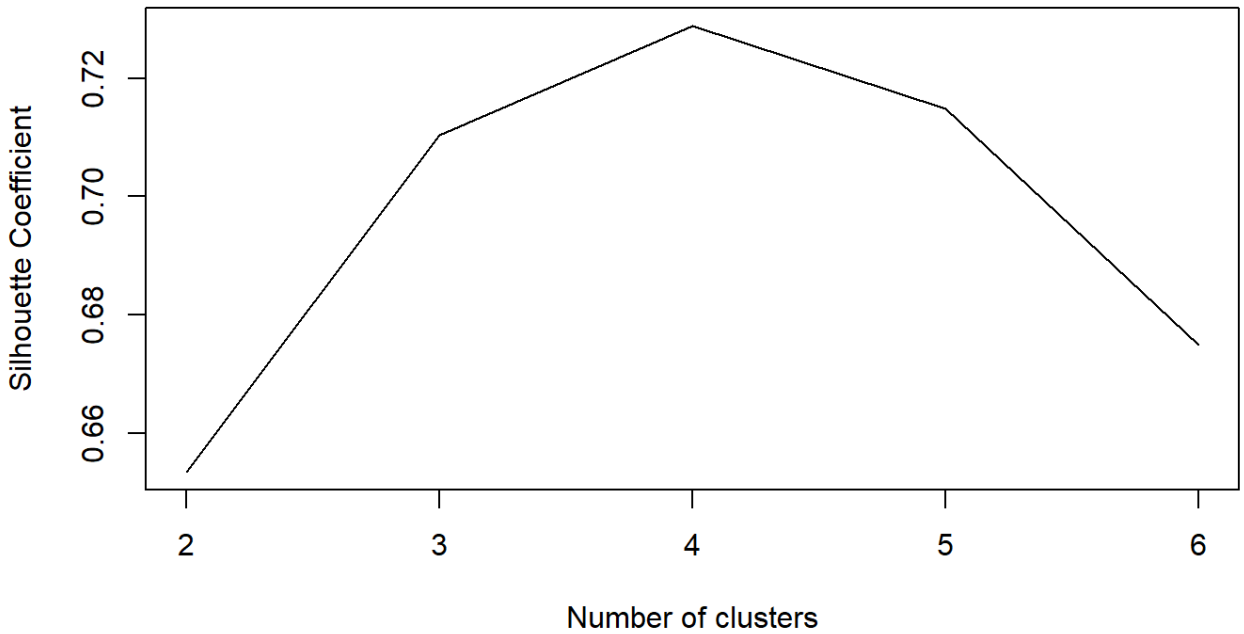
Clustering

- Decision graph of DPC



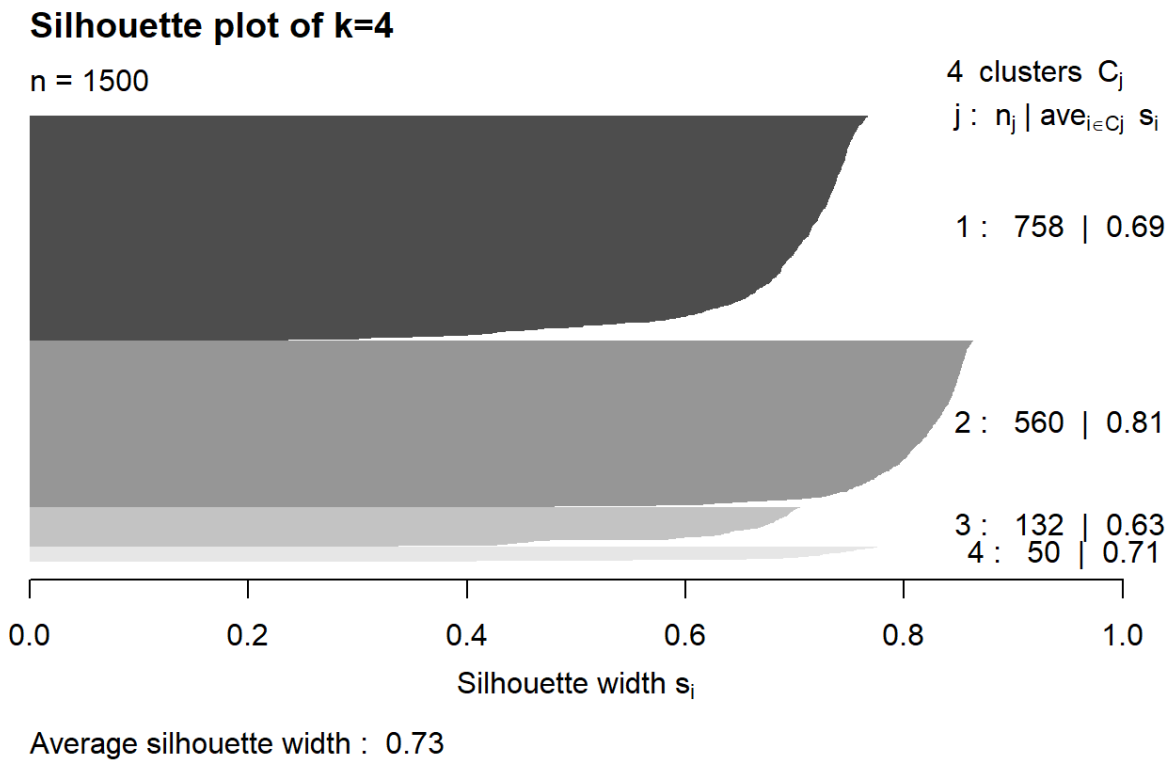
Validation of clustering

- Silhouette coefficient for different number of clusters



Clustering evaluation

- Silhouette plot with four cluster



Clustering results

- Clustering result of DPC