UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Assessment of Resemblance Metrics for Synthetic Data Validation

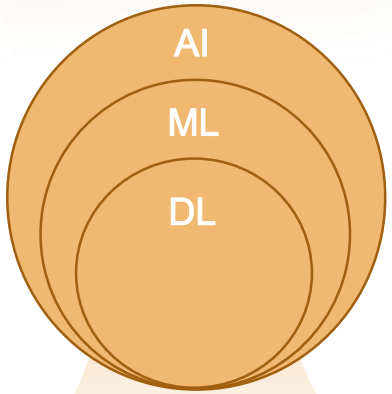
Bachelor's thesis Project
Xinnuo Chen

Advisors: Jordi Cortés Martínez, Daniel Fernández Martínez
Bachelor's degree in Statistics
July 8, 2025

OUTLINE

01 Motivation & Study Purpose

PROBLEM

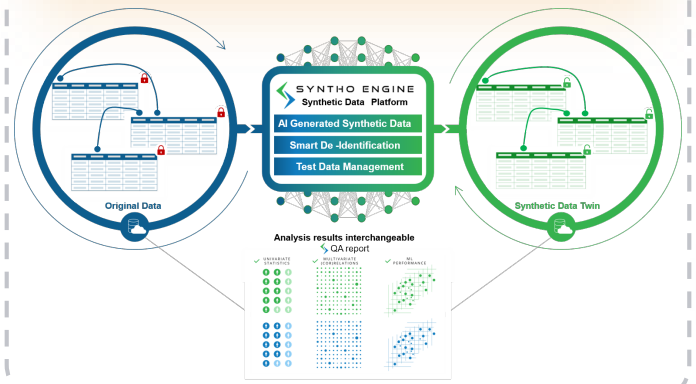


Methods

Large volume of data

POSIBLE SOLUTION

SYNTHETIC DATA



Importance

Data scarcity

Data privacy

Validation

METHODOLOGY

CLUSTERING → K - means

No key metric

Clustering metrics

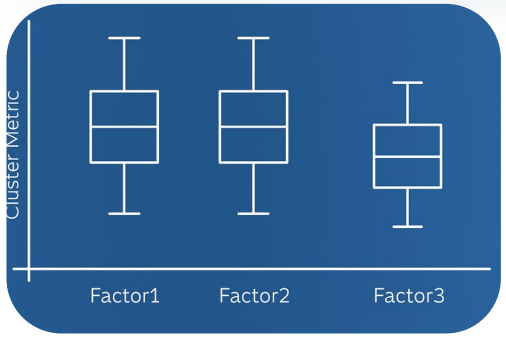
SPECKS

R

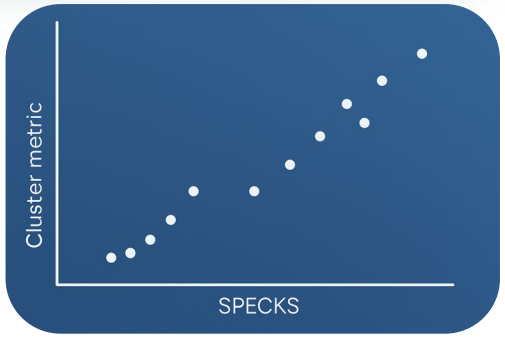
Simulation Study

Purpose

1. Determining factors ?

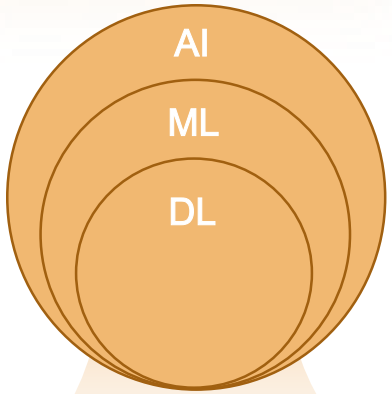


2. Valid indicator ?



01 Motivation & Study Purpose

PROBLEM

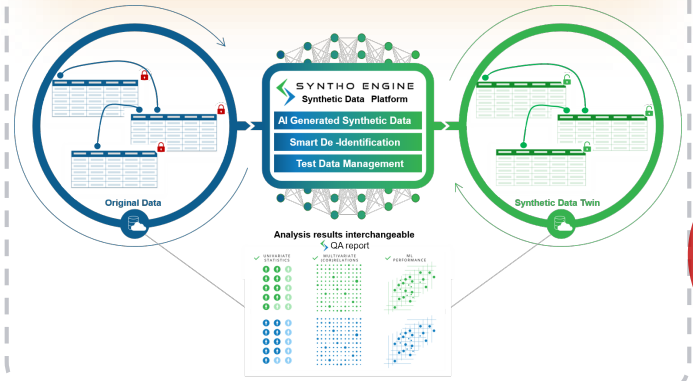


Methods

Large volume of data

POSIBLE SOLUTION

SYNTHETIC DATA



Importance

Data scarcity

Data privacy

Validation

METHODOLOGY

CLUSTERING → K - means

No key metric

Clustering metrics

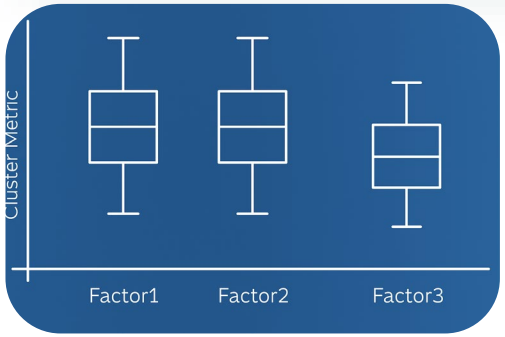
SPECKS

R

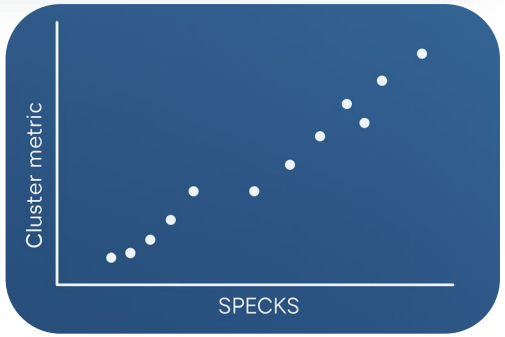
Simulation Study

Purpose

1. Determining factors ?



2. Valid indicator ?



Methodology | Similarity metrics for clustering (1)

Cluster number matching
(Cluster number)

$$p = \frac{\#\ synthetic\ clusters}{\#\ real\ clusters}$$

Optimal value: $p = 1$

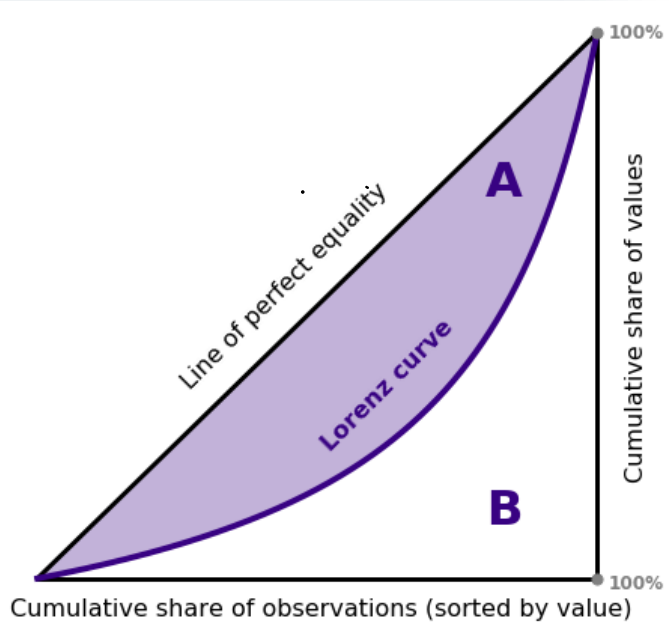
Real data

Synthetic data

Gini Coefficient
(Sample size)

$$G_R = \frac{A_R}{A_R + B_R}$$

$$\Delta G = G_R - G_S$$



Optimal value: $\Delta G = 0$

Methodology | Similarity metrics for clustering (2)

Mean centroid distance
(Central tendency)

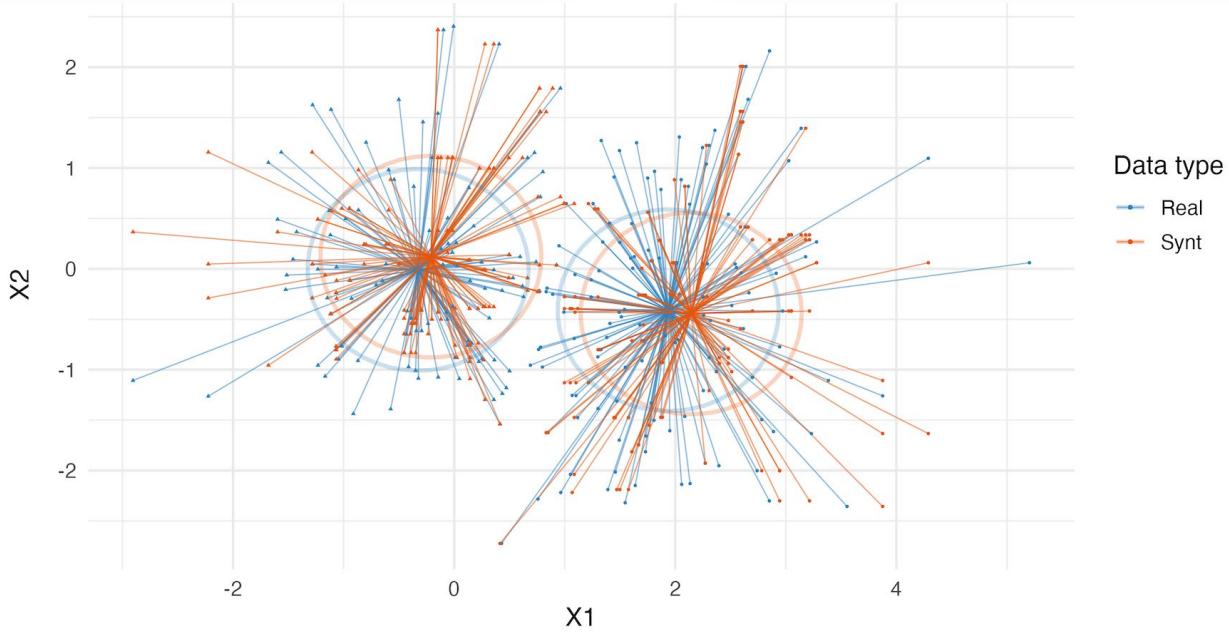
$$D_m(P_i, Q_i) = \frac{1}{K}\sum_{i=1}^{K}\sqrt{(p_{i1} - q_{i1})^2 + (p_{i2} - q_{i2})^2}$$

Optimal value : $D_m(P_i, Q_i) = 0$

Mean variance difference
(Dispersion)

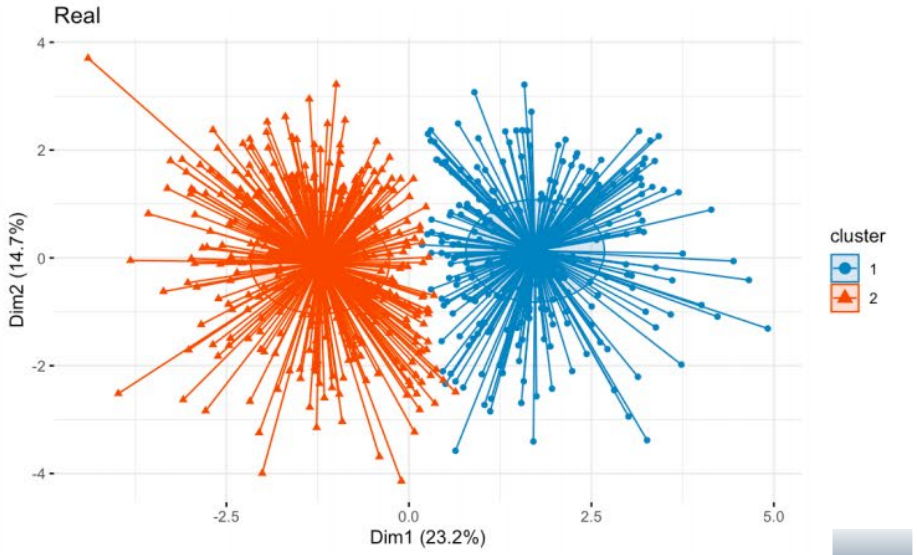
$$D_v(P_i, Q_i) = \frac{1}{K}\sum_{i=1}^{K}|\sigma_{R,i}^2 - \sigma_{S,i}^2|$$

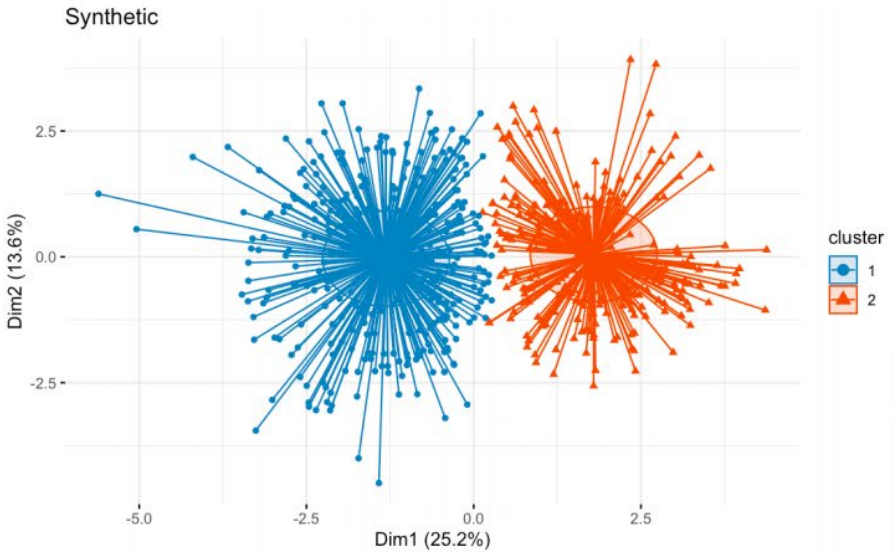
Optimal value : $D_v(P_i, Q_i) = 0$

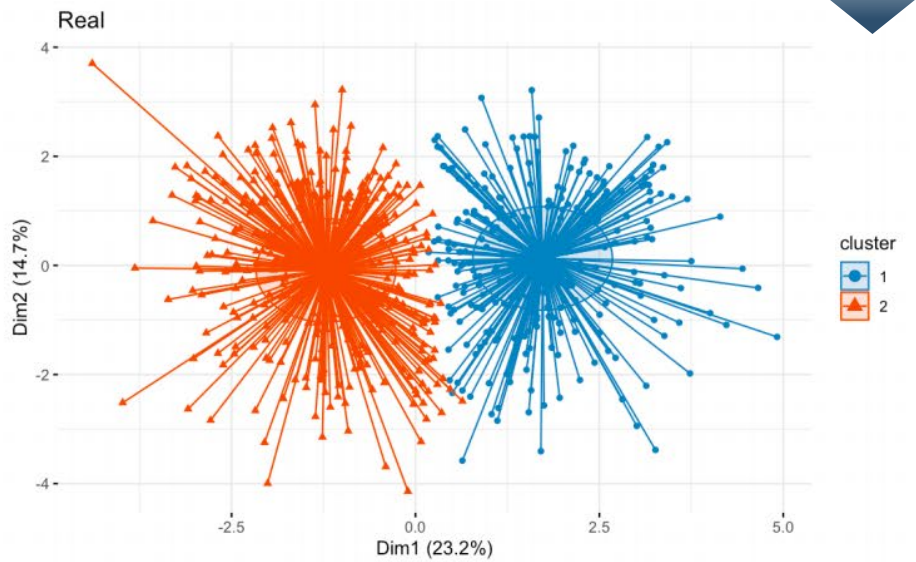


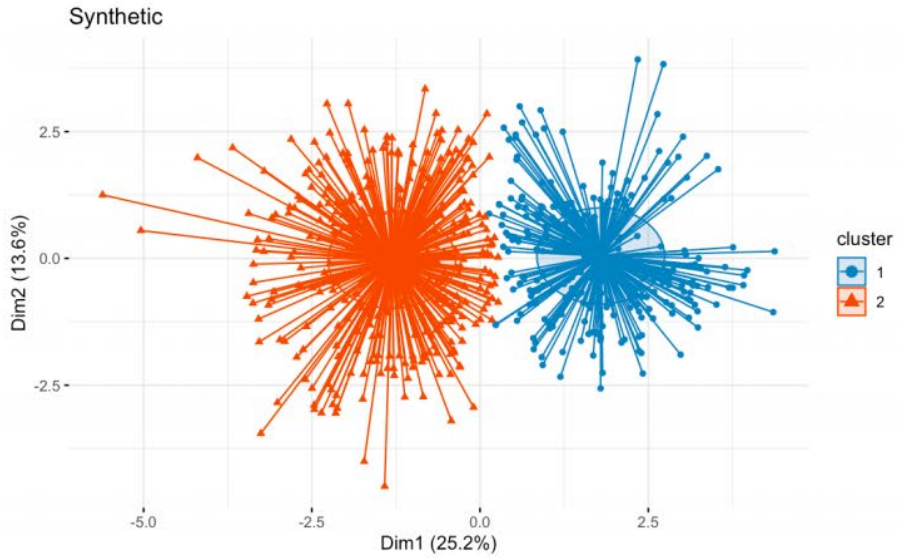
Methodology | Similarity metrics for clustering

Issue : Label Switching (Cluster Matching)





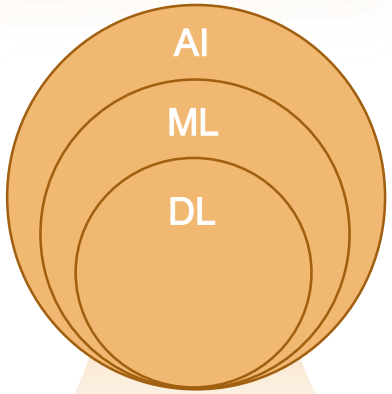




Hungarian algorithm

	Real data cluster	
Synthetic data cluster	1	2
1	3.031	0.473
2	0.349	3.001

PROBLEM

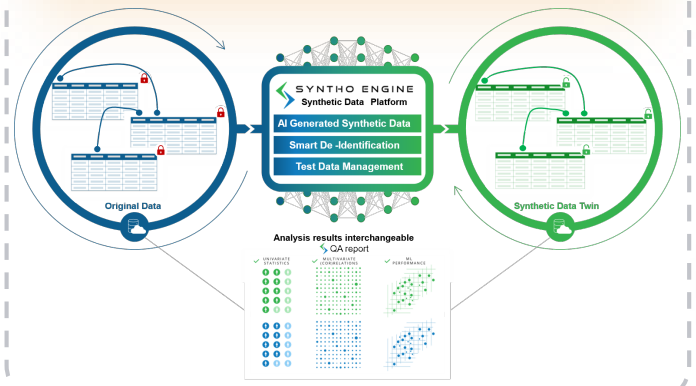


Methods

Large volume of data

POSIBLE SOLUTION

SYNTHETIC DATA



Importance

Data scarcity

Data privacy

Validation

METHODOLOGY

CLUSTERING → K - means

No key metric

Clustering metrics

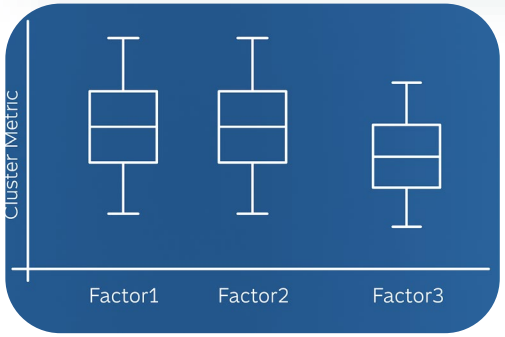
SPECKS

R

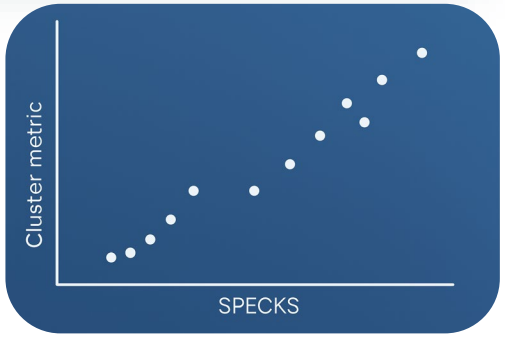
Simulation Study

Purpose

1. Determining factors ?



2. Valid indicator ?



Methodology | Resemblance metric : SPECKS

$$\text{SPECKS} = \sup_{\hat{p}_i} \left| \hat{F}^r(\hat{p}_i) - \hat{F}^s(\hat{p}_i) \right| \overset{H_0}{\sim} KS(n_r, n_s)$$

$$H_0 : F^r(\hat{p}_i) = F^s(\hat{p}_i)$$

$$H_1 : F^r(\hat{p}_i) \neq F^s(\hat{p}_i)$$

Data union

Identification label

V1	V2	Label
160.7181	50.81413	0
159.8003	49.53374	0
160.2217	48.34418	0
…	…	…
180.2575	64.49433	1
180.4691	49.85838	1
181.1581	65.64633	1

Classification Model

Logit
CART
…

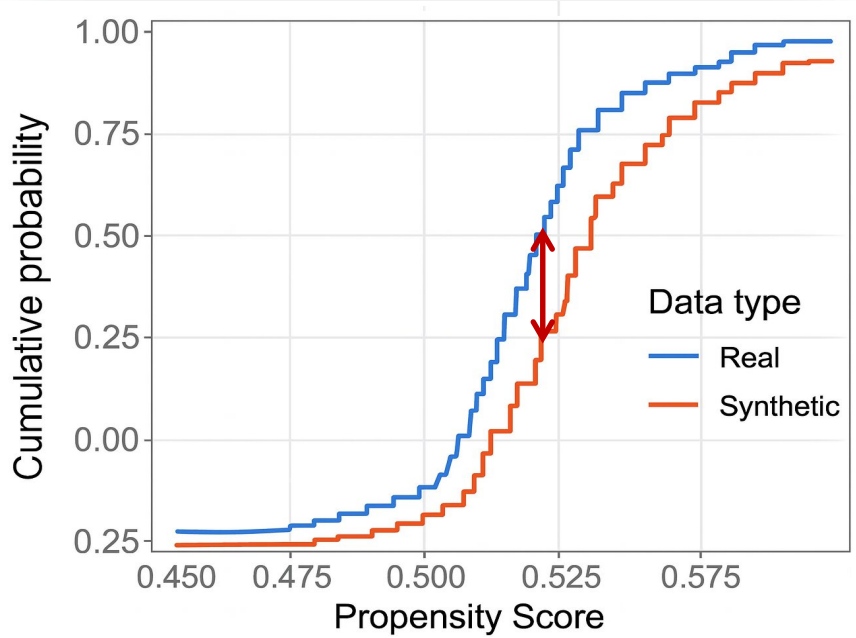
Propensity score

Predicted probability

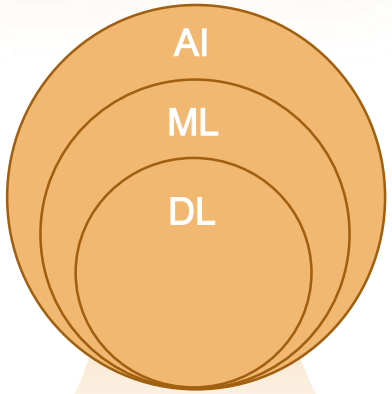
Propensity score
0.2995173
0.2950036
0.3176499
…
0.5864138
0.7624997
0.5929186

Kolmogorov - Smirnov

SPECKS



PROBLEM

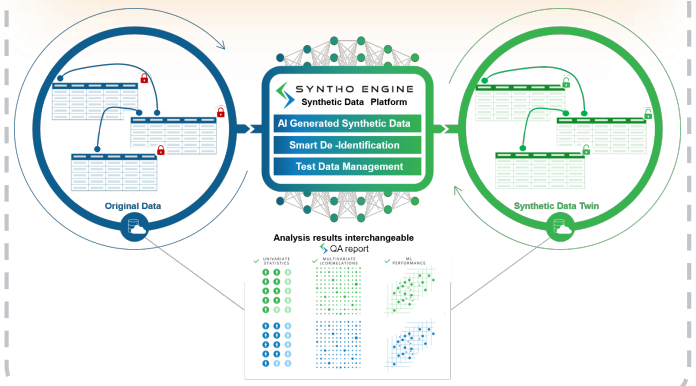


Methods

Large volume of data

POSIBLE SOLUTION

SYNTHETIC DATA



Importance

Data scarcity

Data privacy

Validation

METHODOLOGY

CLUSTERING → K - means

No key metric

Clustering metrics

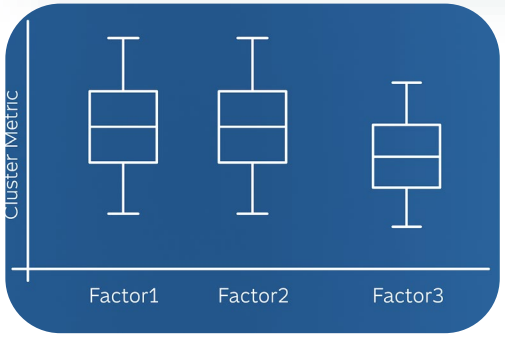
SPECKS

R

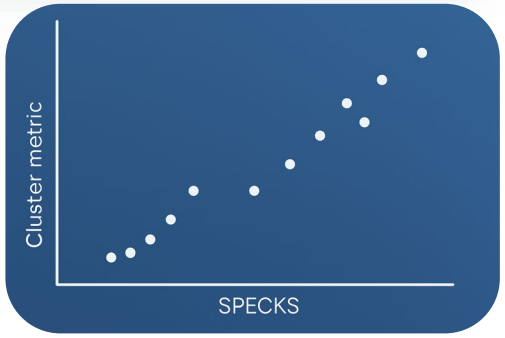
Simulation Study

Purpose

1. Determining factors ?



2. Valid indicator ?



Real data generation

1. Multivariate normal distribution

   R package : *mvtnorm*

2. Setting (factorial design)

Factor	Tested values	Expected quality for higher values
Sample size	50 , 250 , 1000	↑
Number of variables	2, 5, 10	↑
Pearson correlation (ρ)	0, 0.4	↓
Number of clusters	2, 3, 4	↓
Cluster separation	0σ, 2σ, 6σ, 10σ	↓

Synthetic data generation

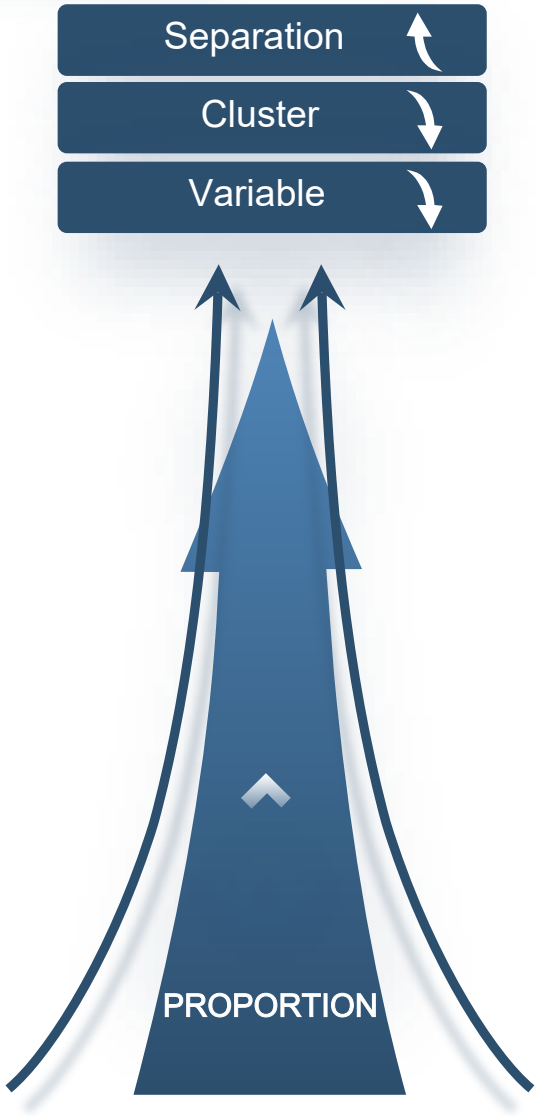
1. CART generation method

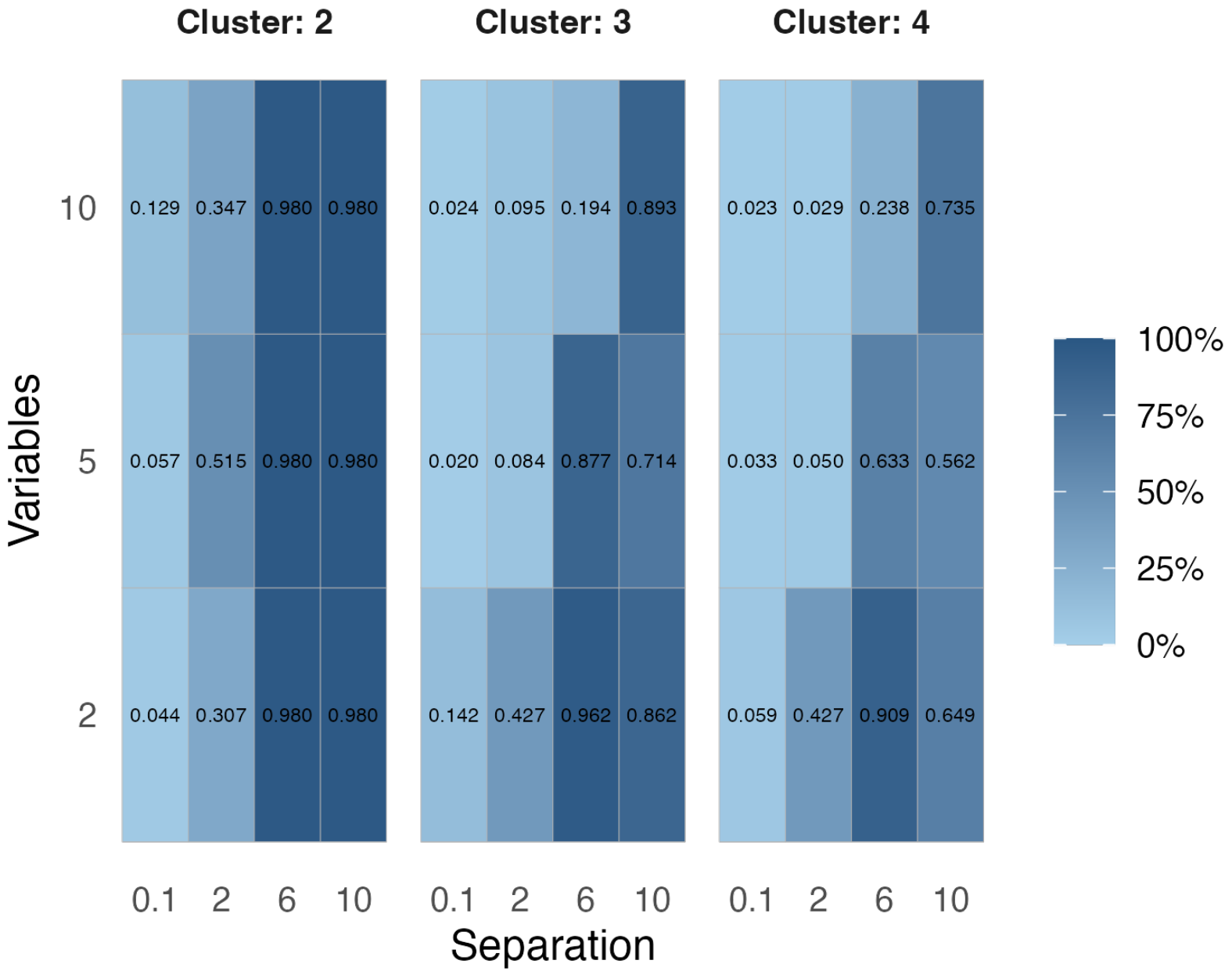
   R package: *synthpop*

2. Computacional cost

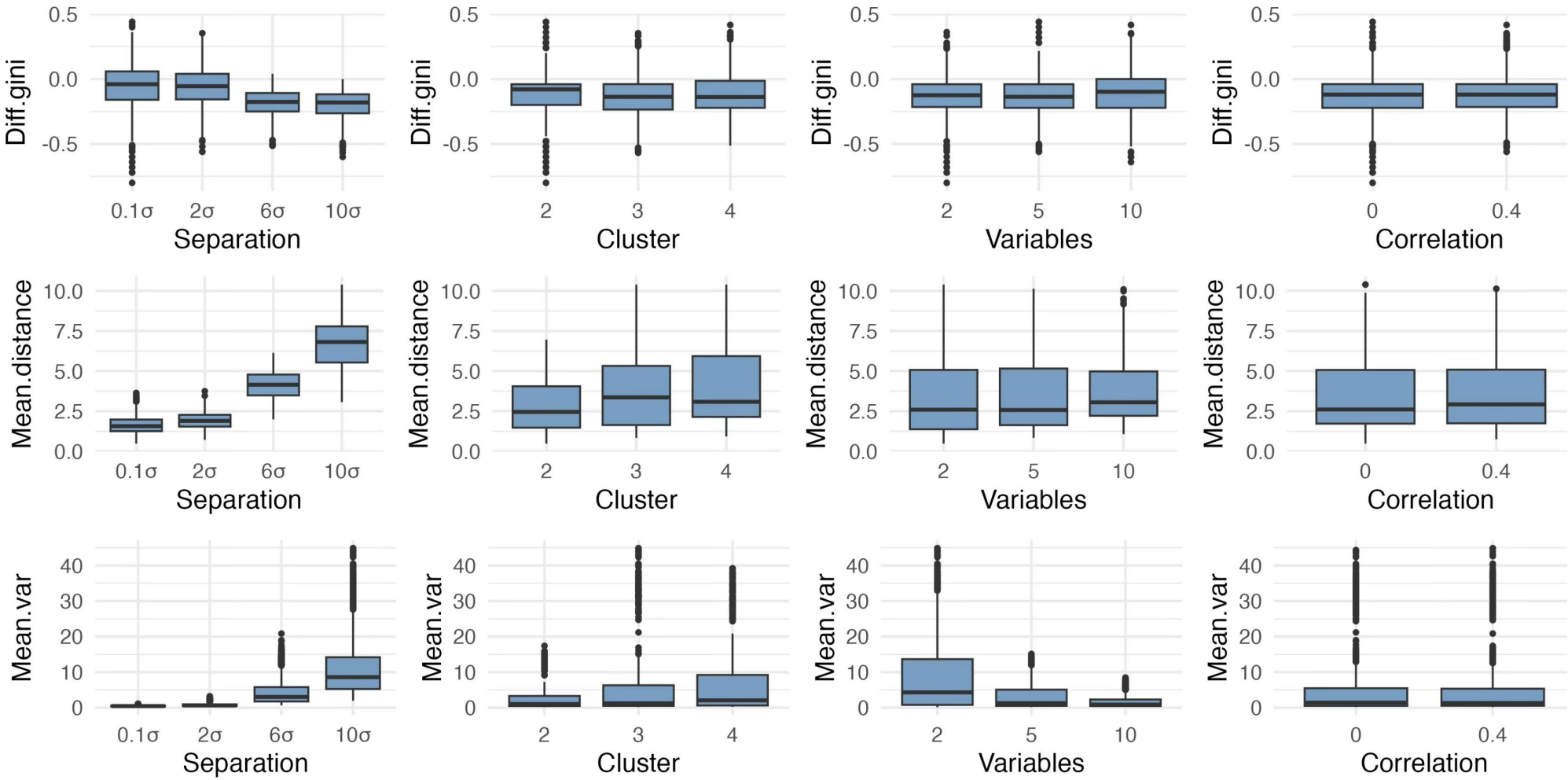
Method	User	System	Elapsed
cart	0.627	0.635	255.221
norm	1.050	0.953	325.169

Results | Cluster number matching proportion





Results | Clustering structure comparison

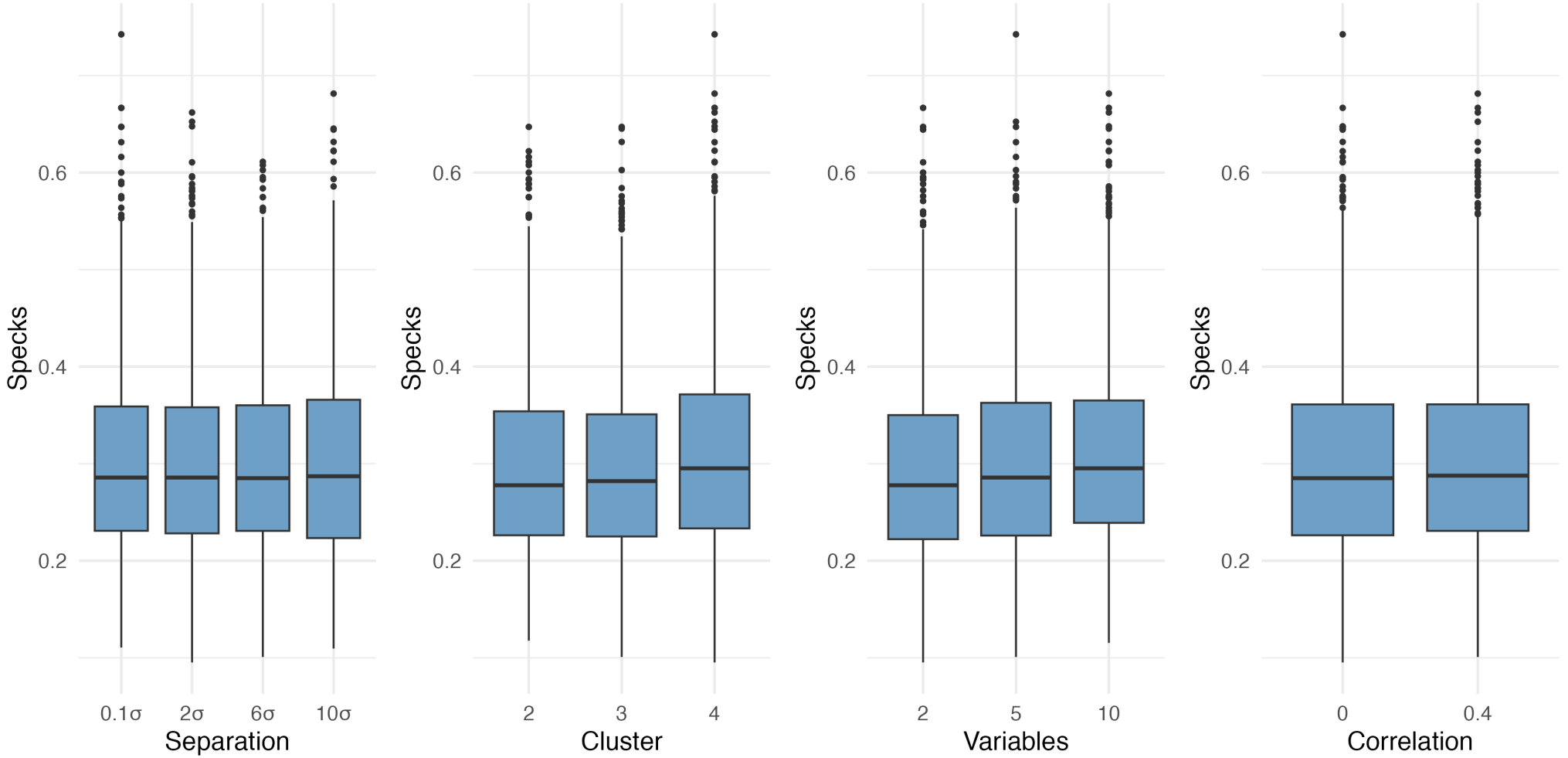


Better maintained structure :

- Separation
- Cluster
+ Variable

Results | SPECKS Behaviour

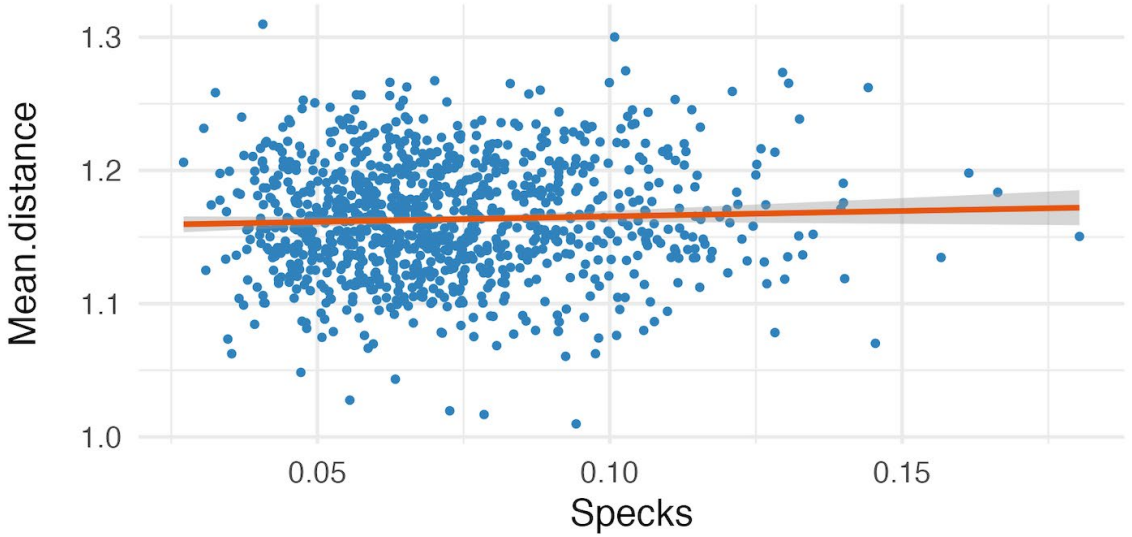
Insentitive to clustering structure changes



Results | Ideal vs. Problematic Scenarios

Ideal Scenario : Better structure preservation

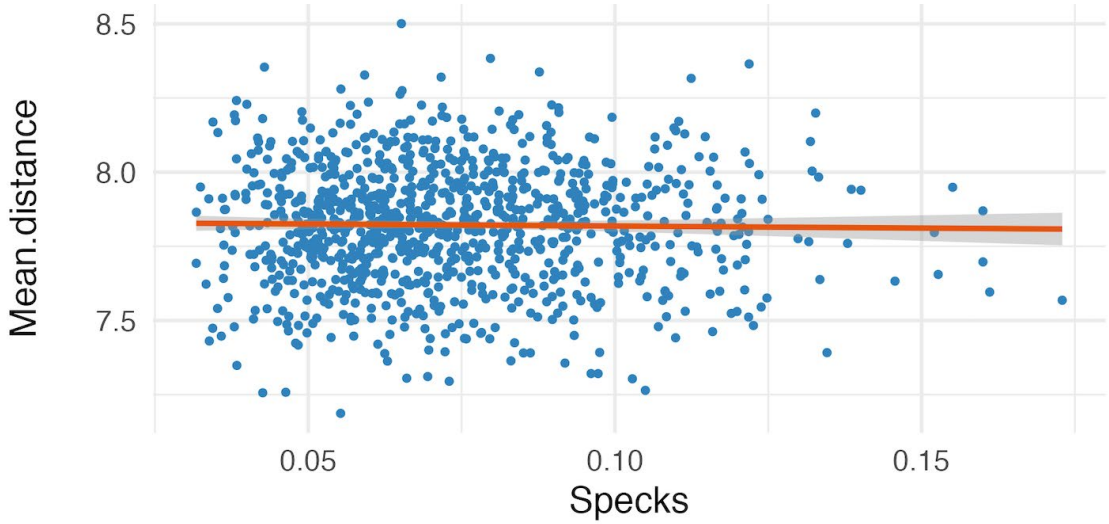
- 2 clusters
- 2σ separation



VS

Problematic Scenario : Poorer structure preservation

- 4 clusters
- 10σ separation



Conclusion & Future research direction

PURPOSE 1

Datasets with fewer clusters and moderate separation, leads to a more similar cluster structure

PURPOSE 2

SPECKS is not suitable for evaluating structural quality in cluster analysis

- Test alternative metrics
- Explore non -normal distributions
- Explore other clustering methods and other statistical methods
- Evaluate larger sample sizes

05 References

Nowok, B., Raab, G. M., and Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26.

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-Dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., et al. (2023). A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46.

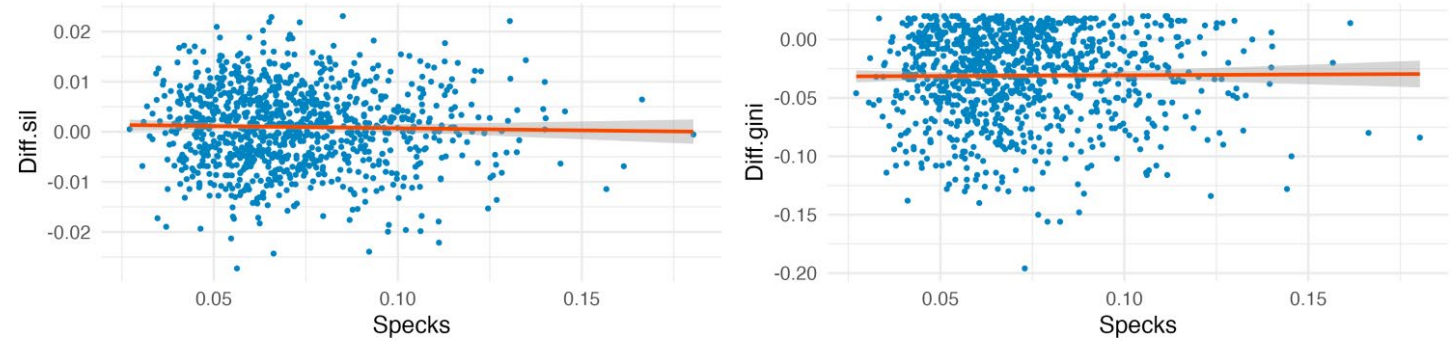
Raghunathan, T. E. (2021). Synthetic data. *Annual review of statistics and its application*, 8(1):129–140

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3):663–688.

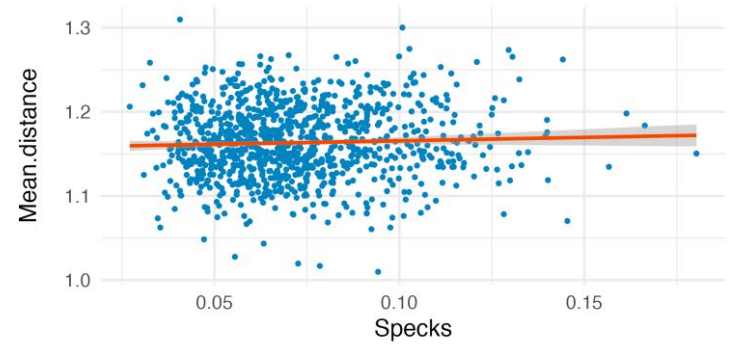
Mills-Tettey, G. A., Stentz, A., and Dias, M. B. (2007). The dynamic hungarian algorithm for the assignment problem with changing costs. *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*.

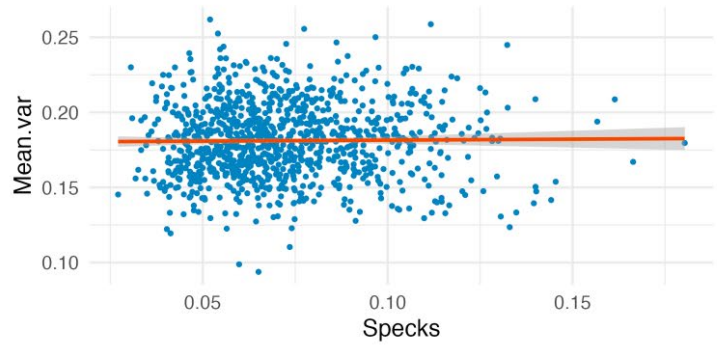
THANKS FOR YOUR ATTENTION

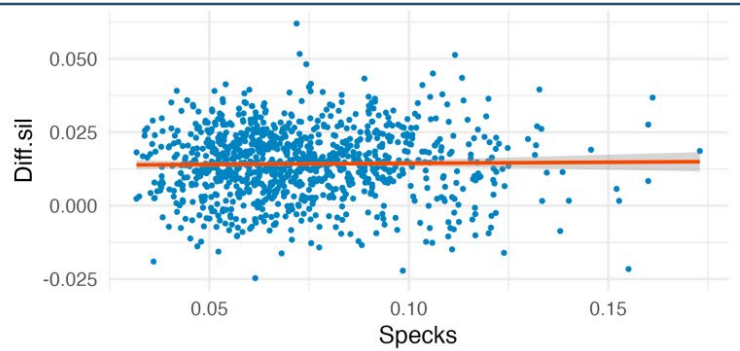
R code

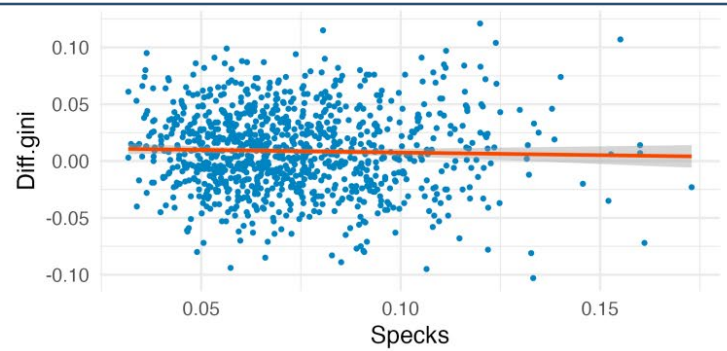

Diff.sil
0.02
0.01
0.00
-0.01
-0.02
0.05
0.10
0.15
Specks

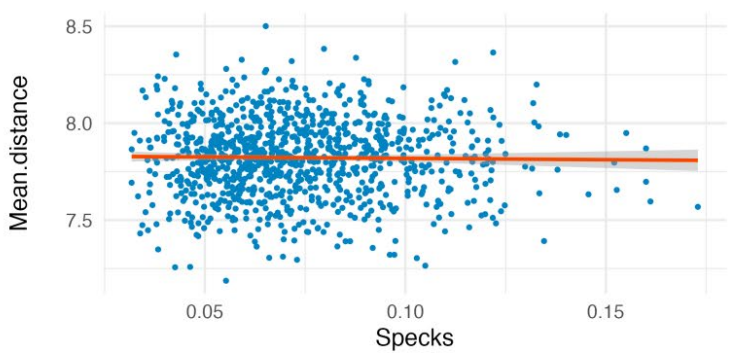
Diff.gini
0.00
-0.05
-0.10
-0.15
-0.20
0.05
0.10
0.15
Specks

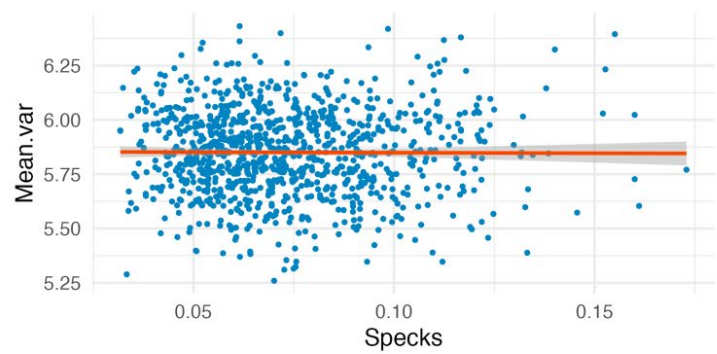

Mean.distance
1.3
1.2
1.1
1.0
0.05
0.10
0.15
Specks

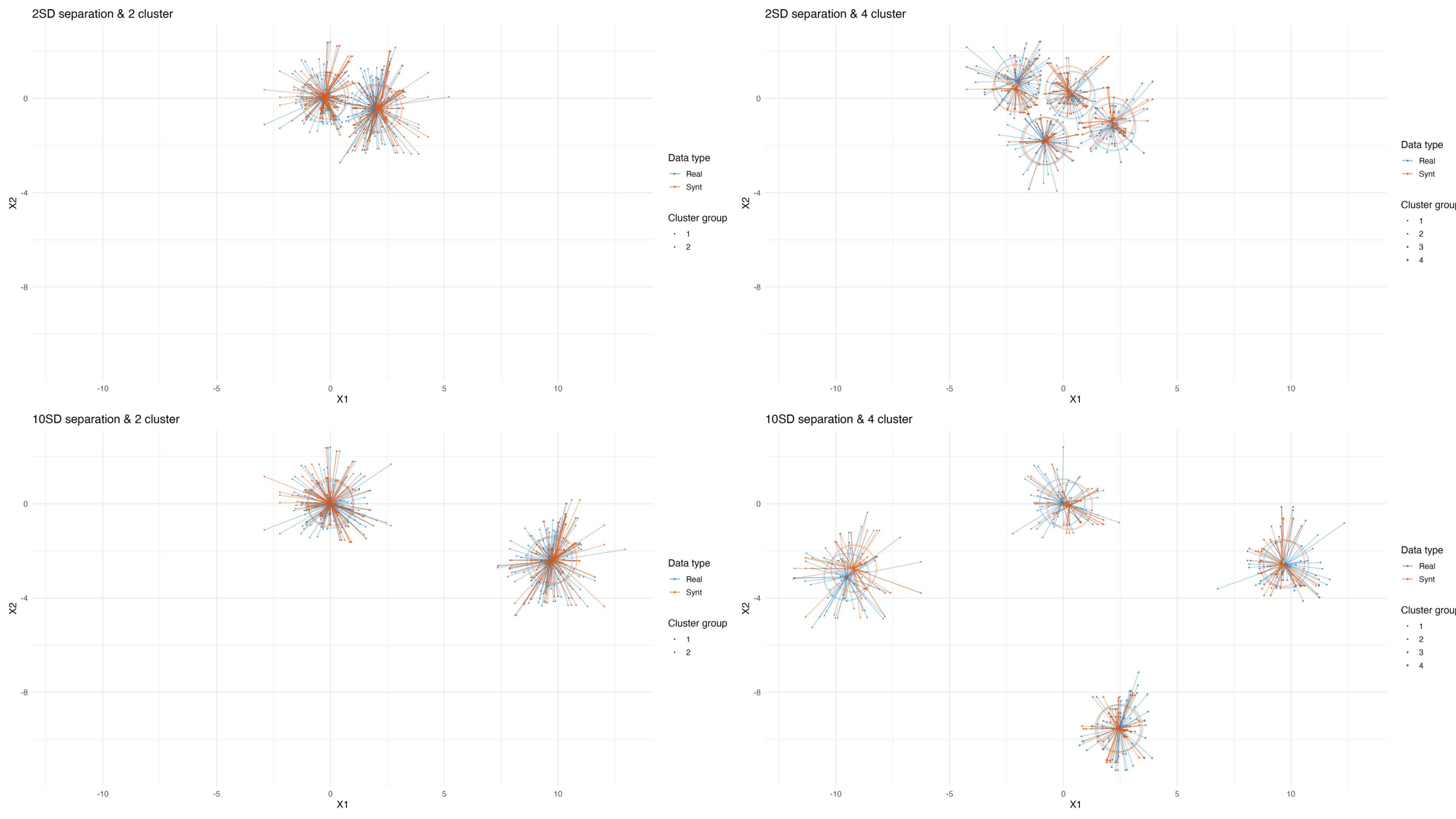

Mean.var
0.25
0.20
0.15
0.10
0.05
0.10
0.15
Specks

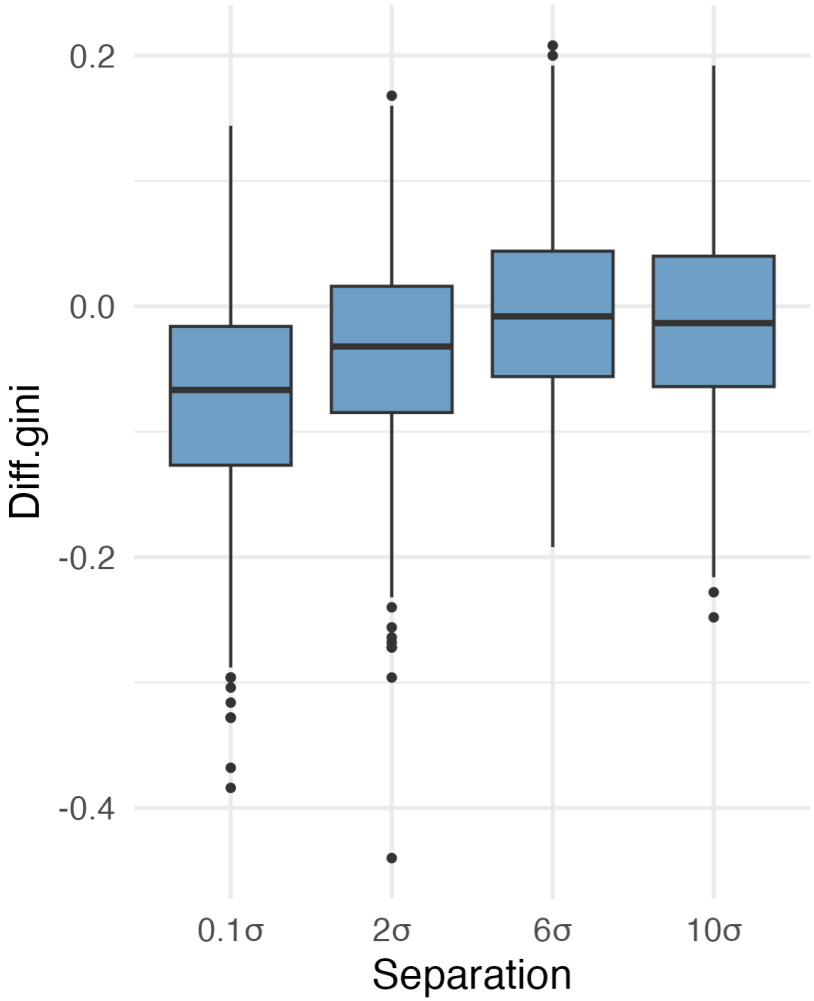

Diff.sil
0.050
0.025
0.000
-0.025
0.05
0.10
0.15
Specks

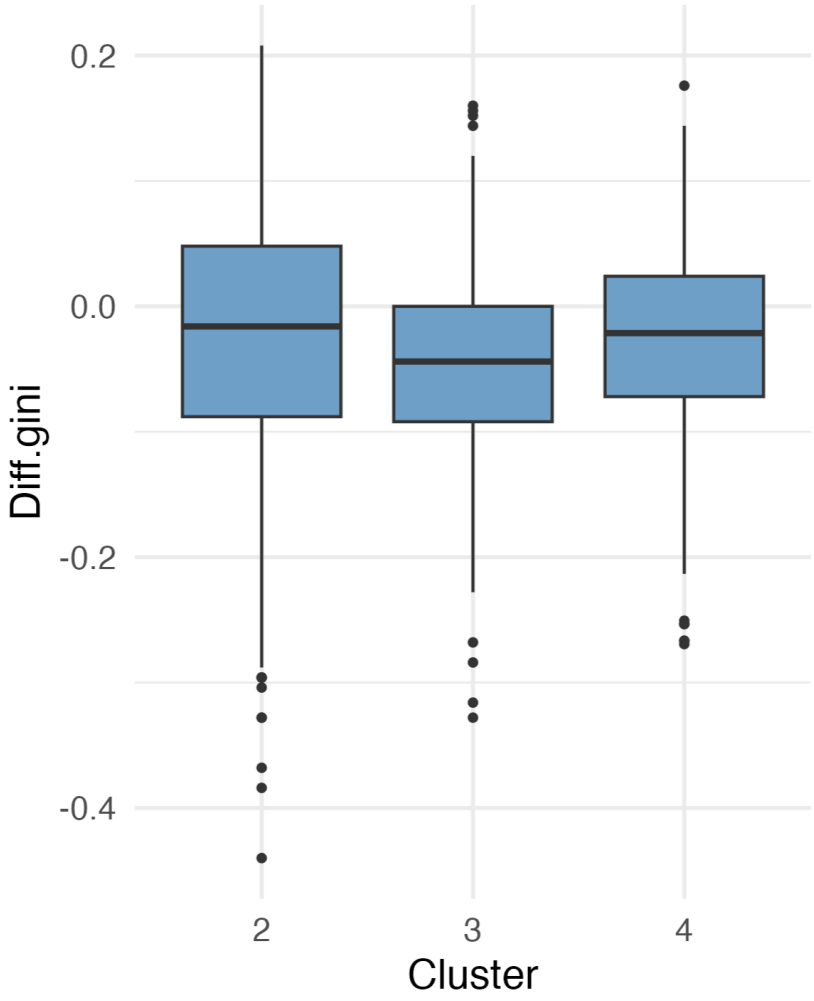

Diff.gini
0.10
0.05
0.00
-0.05
-0.10
0.05
0.10
0.15
Specks

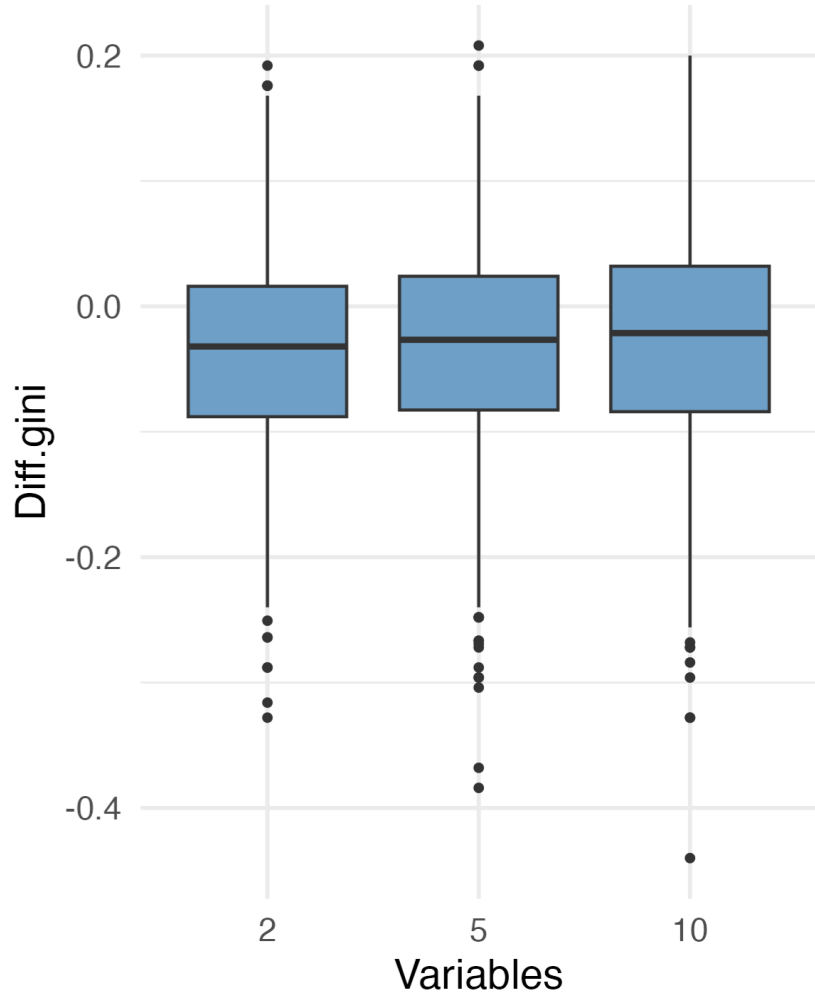

Mean.distance
8.5
8.0
7.5
0.05
0.10
0.15
Specks

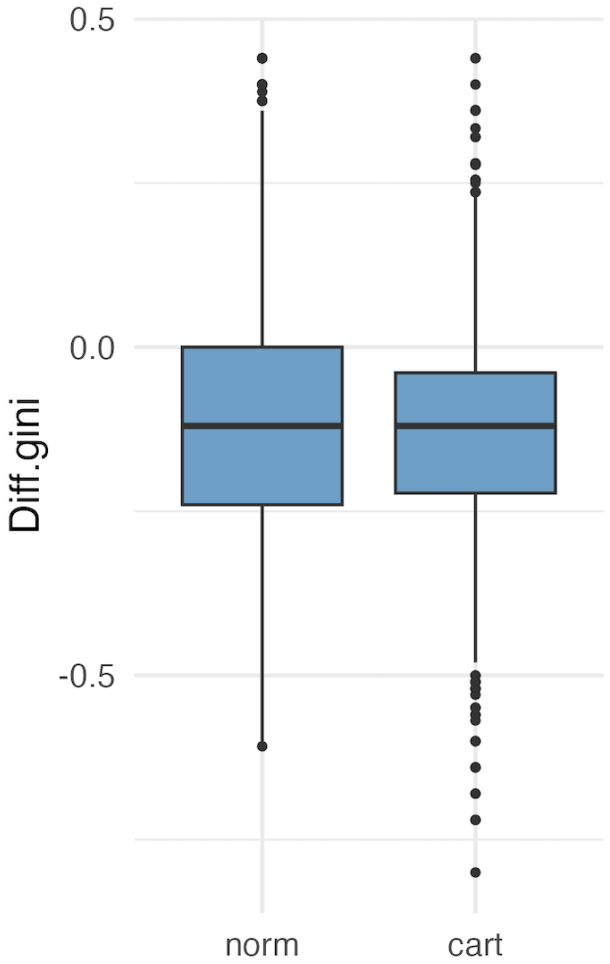

Mean.var
6.25
6.00
5.75
5.50
5.25
0.05
0.10
0.15
Specks

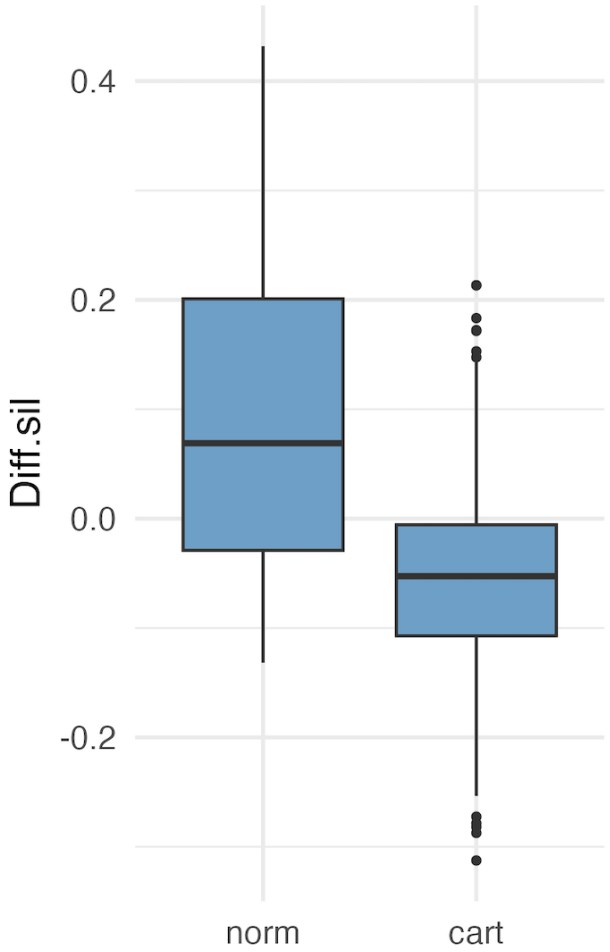

2SD separation & 2 cluster
2SD separation & 4 cluster
10SD separation & 2 cluster
10SD separation & 4 cluster
X1
X2
Data type
Real
Synt
Cluster group
1
2
3
4

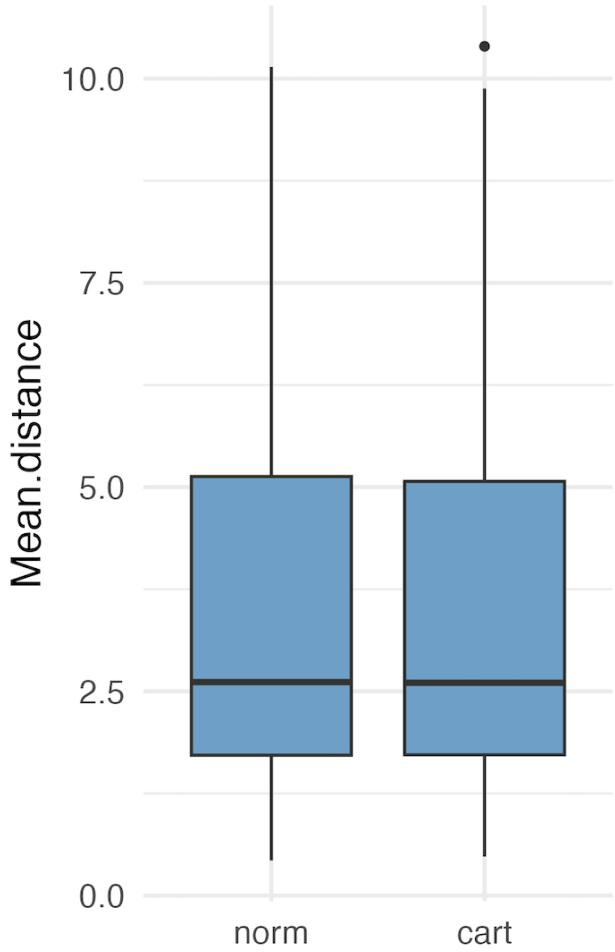

0.2
0.0
-0.2
-0.4
Diff.gini
0.1σ
2σ
6σ
10σ
Separation

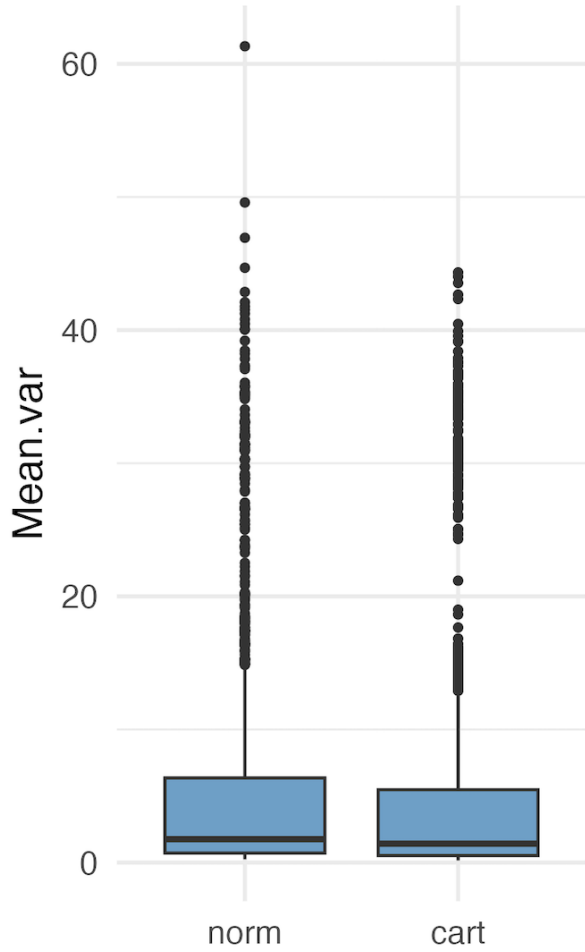

0.2
0.0
-0.2
-0.4
Diff.gini
2
3
4
Cluster


0.2
0.0
-0.2
-0.4
Diff.gini
2
5
10
Variables


0.5
0.0
-0.5
Diff.gini
norm
cart


0.4
0.2
0.0
-0.2
Diff.sil
norm
cart


10.0
7.5
5.0
2.5
0.0
Mean.distance
norm
cart


60
40
20
0
Mean.var
norm
cart