BACHELOR'S DEGREE THESIS

Degree in Statistics

Title: Assessment of the Resemblance Metrics for Synthetic data validation

Author: Xinnuo Chen

Advisor: Jordi Cortés Martínez, Daniel Fernández Martínez

Department: Statistics and Operations Research (Universitat Politècnica de Catalunya-BarcelonaTECH)

Academic year: 2024 - 2025

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Abstract

In the context of the constant growth of artificial intelligence, the requirement for large volumes of data has become one of the main challenges. Using synthetic data is a viable alternative for addressing both the scarcity of real data and the need to protect information privacy. For synthetic data to be useful, it is essential to validate that the characteristics of the original data are preserved.

This study analyses the reliability of the SPECKS metric for measuring similarity between real and synthetic data in cluster analysis. Several factors affecting the ability of algorithms to replicate the structure of the original clusters are examined through simulations. The relationship between SPECKS and clustering metrics that allow the similarity of the clusters' structure to be evaluated is also studied to determine whether SPECKS can be a good indicator of the quality of structural preservation in synthetic data clusters.

The results suggest that SPECKS is insensitive to structural changes and is therefore not a suitable metric for evaluating structural quality in cluster analysis.

Keywords: Synthetic data; SPECKS; Clustering; K-means; Clustering metrics; Simulation

Resum

En un context de creixement constant de la intel·ligència artificial, la necessitat de grans volums de dades s'ha convertit en un dels principals reptes. L'ús de les dades sintètiques és una alternativa viable per fer front a l'escassetat de dades reals i la necessitat de protegir la privacitat de la informació. Perquè aquestes dades siguin útils, la validació de la conservació de les característiques de les dades originals és imprescindible.

Aquest estudi analitza la fiabilitat de la mètrica de semblança SPECKS per mesurar la similitud entre dades reals i sintètiques en anàlisis de clústers. Mitjançant simulacions, s'examinen diversos factors per analitzar com afecten en la capacitat dels algoritmes per replicar l'estructura dels clústers originals. També s'estudia la relació de SPECKS amb mètriques de clusterització per avaluar la similitud de l'estructura dels clústers obtinguts, per determinar si és un bon indicador de la qualitat de preservació estructural dels clústers de les dades sintètiques.

Els resultats indiquen que SPECKS no és sensible als canvis estructurals i, per tant, no es considera una mètrica adequada per avaluar la qualitat estructural en l'anàlisi de clústers.

Paraules claus: Dades sintètiques; SPECKS; Anàlisi de clústers; K-means; Mètriques de clusterització; Simulació

AMS Classification

- 62-07 Data analysis
- 62H30 Classification and discrimination; cluster analysis
- 65C99 None of the above, but in this section

Acknowledgements

I would like to express my sincere gratitude to my tutors, Jordi Cortés and Daniel Fernández, for giving me the opportunity to develop this research topic, as well as for their constant support and guidance throughout my final project. Their involvement, availability, sense of responsibility, and interest were all crucial to the completion of this work. I am also very grateful to Nora Amama for her assistance during the whole process. Not only for her technical expertise and professionalism, but also for the personal insight and advice she shared from her own experience.

I would also like to thank my family and friends for their unconditional support and encouragement.

Finally, I extend my appreciation to the university and to all the professors over the course of my degree. I have been very fortunate to have been taught by an excellent, passionate, and dedicated team of teachers who have provided me with the necessary tools to grow academically and professionally.

Contents

List of Figures vi

List of Tables viii

1 Introduction 1

1.1 Motivation & Background 1

1.2 Purpose 2

1.3 Structure of the document 3

2 Methodology 4

2.1 Clustering 4

2.1.1 Clustering method 4

2.1.2 Cluster comparative metrics 8

2.1.3 Label switching 11

2.2 Synthetic data 13

2.2.1 Synthetic data generation method 13

2.2.2 Resemblance metric 14

2.3 Tools for assessing SPECKS-variable relationship 16

2.4 Simulation 17

2.4.1 Original data 18

2.4.2 Synthetic data 20

2.5 Software 21

3 Results 22

3.1 Influence of Factors on the Clustering Quality of Synthetic Data 22

3.1.1 Matching number of clusters analysis 22

3.1.2 Factors behavior according to clustering metrics in different scenarios . 23

3.1.3 Gini behaviour for unbalanced cluster 28

3.1.4 Cluster centroid behaviour 29

3.2 SPECKS: Using a Resemblance Metric as a Proxy for Clustering Utility 30
3.2.1 Factors 30
3.2.2 Scenarios for further analysis 32

4 Discussion **35**

5 Appendix **41**
5.1 Supplementary Figures 41
5.2 Code 42

List of Figures

1.1 The relationship between artificial intelligence, machine learning and deep learning . . . 1

1.2 Synthetic data generation and its analytical utility representation . . . 2

2.1 Cluster assignment visual representation . . . 4

2.2 Illustration of hierarchical clustering process and dendrogram output . . . 5

2.3 K-means clustering applied to a two-cluster scenario . . . 6

2.4 Elbow method for determining optimal number of clusters . . . 7

2.5 Illustration of Silhouette coefficient calculation . . . 7

2.6 Representation of centroid distances between real and synthetic data . . . 10

2.7 Cluster assignment before label switching correction . . . 11

2.8 Cluster assignment after label switching correction . . . 12

2.9 Cumulative distribution of Propensity Scores . . . 15

2.10 Representation of cluster separation based on the distance between their respective centroids, expressed in terms of standard deviations σ. . . . 19

3.1 Proportion of matching cluster number between real and synthetic data, across different separations, number of variables, and number of clusters . . . 23

3.2 Comparison of Clustering Quality Metrics by Synthetic Data Generation Method (*CART & Norm*) . . . 24

3.3 Effect of Cluster Separation, Number of Clusters, and Variables on *Diff.sil* by Generation Method . . . 25

3.4 Distribution of clustering metrics (*Diff.gini*, *Mean.distance*, *Mean.var*) according to four factors: cluster separation, number of clusters, number of variables, and variable correlation. . . . 26

3.5 *Diff.gini* values across different levels of separation, number of clusters, and number of variables in unbalanced cluster scenarios. The x-axis represents the varying factor; the y-axis shows the difference in Gini index between real and synthetic data . . . 28

3.6 Comparison of Real and Synthetic Cluster Centroids . . . 29

3.7 SPECKS values across different levels of separation, number of clusters, and number of variables. The x-axis shows the varying factor; the y-axis represents the SPECKS metric. 30

3.8 SPECKS values (x-axis) across and the difference in Gini index (y-axis) under three different cluster matching scenarios. 31

3.9 SPECKS values (x-axis) across and the difference in Silhouette index (y-axis). . 32

3.10 SPECKS values (x-axis) across and the difference in clusters mean distance (y-axis). 32

3.11 SPECKS values (x-axis) across and the difference in clusters mean variance (y-axis). 32

3.12 Relationship between SPECKS and structural difference metrics in a scenario with two moderately separated clusters (2σ), chosen for its stability and expected structure preservation. 33

3.13 Relationship between SPECKS and structural difference metrics in a scenario with high separation (10σ) and four clusters, chosen for conditions known to hinder structure preservation in synthetic data. 34

5.1 Comparison of Cluster Centroids Between Real and Synthetic Data Across Varying Cluster Counts and Separation Levels 41

List of Tables

2.1 Distance between centroids 12

2.2 Dataset on the left containing two variables and the true label (real or synthetic). Dataset on the right containing the same two variables and the propensity scores. 14

2.3 Summary of the experimental design factors, the specific values tested for each, and the expected influence they may have on the resemblance between real and synthetic data. 19

2.4 R packages used and their purpose in the study 21

3.1 Execution time for 1 iteration 26

5.1 R packages used and their purpose in the study 42

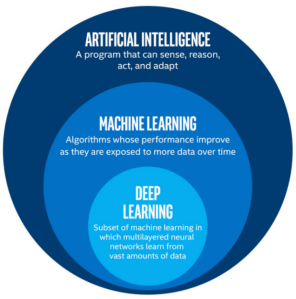
1 Introduction

1.1 Motivation & Background

In recent years, the concept of artificial intelligence has gained considerable traction and visibility in various societal domains, including health, education, industry, economy, among others (Polak and Anshari, 2024). A representative illustration of this phenomenon is the advent of ChatGPT (OpenAI, 2024), a tool that has exerted a substantial influence on both the technological realm and quotidian life, a tool that has had a considerable impact on both technological and quotidian life, changing the way people access information, generate content, and interact with technology.

Concepts such as artificial intelligence, machine learning deep learning, etc., are constantly evolving and play a fundamental role in technological progress. As shown in *Figure 1.1*, these concepts are hierarchically related: deep learning is a branch of machine learning, which is in turn a branch of artificial intelligence. This continuous evolution is fostered by the training of algorithms and models that require large volumes of data, and the insufficiency of these has become one of the most pressing challenges within the sector (Alzubaidi et al., 2023). In this context, the generation of synthetic data offers a potential solution to the problem of real data scarcity. These data sets are of an artificial nature and are derived from real-world data sets. The aim is to replicate the structure, distribution, statistical properties, configuration, and other parameters of the original data sets, while maintaining a completely fictitious nature.

Figure 1.1: The relationship between artificial intelligence, machine learning and deep learning

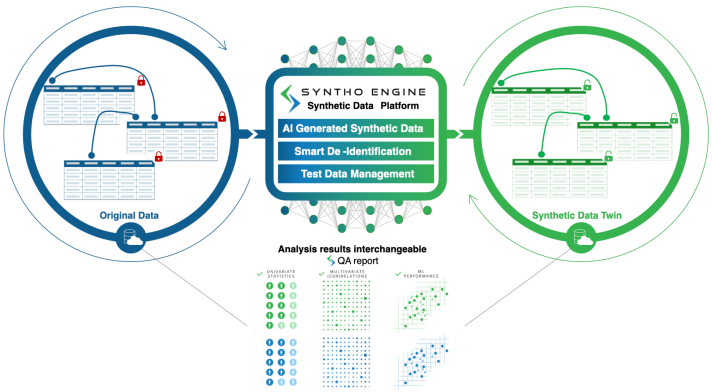


The analysis of data and the information derived from it provides valuable insights, and its relevance and value have grown steadily as a result, forming an essential part of progress and evolution. However, despite the availability of valuable databases, data protection and privacy issues often limit their use. As well as providing a solution to the lack of real data, one of the main applications of synthetic data is protecting information. Medical, banking, educational and governmental databases, among others, may contain sensitive personal information such

as names, identity documents and bank account details, which makes them difficult to use for analysis due to the risk of disclosure. Using synthetic data makes it possible to anonymize these records. No record in the generated set corresponds to a specific real-world record, thus enabling the reproduction of characteristics and behavior while preserving privacy and avoiding the exposure of any individual's sensitive or personal information. Information obtained from synthetic data should be indistinguishable from that which would be extracted from real data while maintaining the privacy of sensitive information.

Synthetic data is becoming increasingly popular due to its ability to solve some of the sector's most pressing problems. It is already used in a wide range of fields, including energy, education, healthcare and finance, among others. While extensive algorithms for generating synthetic data already exist in the literature and industry, generation is not the only critical point — usability is also essential. The quality of a synthetic dataset depends on its usability. Generated data is only considered valid if it retains similar statistical properties as the real dataset it replicates. *Figure 1.2* illustrates how synthetic data maintains analytical equivalence between real and synthetic data while preserving privacy and enabling accessibility without revealing personal information. To validate usability, it is necessary to verify that these artificial records correctly reproduce the characteristics, structure and behavior of the original data and that analyses lead to the same conclusions in both sets.

Figure 1.2: Synthetic data generation and its analytical utility representation



This study evaluates a resemblance metric that compares synthetic and real data. The aim is to assess its usefulness as an indicator of whether the synthetic data can support meaningful cluster analysis.

1.2 Purpose

This final degree project presents a simulation study that aims to evaluate the reliability of the Specks assembly metric in reflecting the quality of synthetic data for clustering analysis. Details

of the metric and the analysis will be provided in the methodology section. To achieve this, the study is divided into two sub-objectives:

1. **Factors influencing clustering similarity**: A preliminary study was conducted to determine whether factors such as the number of variables, sample size, degree of correlation of the real data sets, or the method used to generate the synthetic data could affect the similarity between clustering results obtained from real and synthetic data. If any such factors are identified, guidelines can be developed for selecting the most appropriate real dataset for clustering analysis based on a synthetic dataset.

2. **Predictive capability of Specks**: The analysis examines the extent to which the Specks value can anticipate the concordance of real and synthetic cluster patterns. The aim is to quantify the similarity between the original and synthetic data using the Specks statistic and determine whether this value is related to the clustering metrics that describe the similarity of the cluster patterns obtained in both sets. If a relationship is proven, a simple assembly measure could predict how well the synthetic data preserves the cluster patterns of interest. This would enable informed decisions to be made in advance regarding the suitability of the synthetic data for the intended analysis.

1.3 Structure of the document

The introduction mainly describes the motivation and objective of the work. It also illustrates the memory's overall structure and the sections that comprise it. The methodology section presents the clustering metrics used to compare the real and synthetic cluster results, and offers an in-depth overview of the Specks assembly statistic. It also describes the different study scenarios, the factors employed in each simulation, and the most relevant software packages. The results section presents the findings obtained in relation to each objective. These findings are supported by graphics and tables. Finally, the Discussion section highlights and interprets the most significant results. It also emphasizes the limitations encountered during the development of the work. There is also a brief section that compares the results of previous studies on the subject. The study's conclusions and the references used to develop them are presented last.

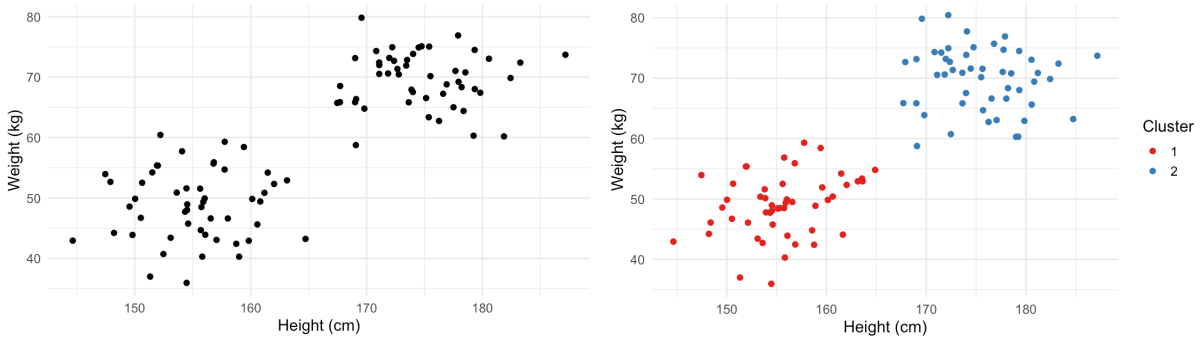
2 Methodology

2.1 Clustering

Clustering is a machine learning algorithm that analyses data structures and aims to organize observations from a dataset into smaller groups according to their characteristics. The resulting groupings present similar patterns within groups and differences between groups. These groupings are called clusters. The objective of this process is to maximize the internal cohesion of the clusters, to make the observations within each group as similar as possible, and to minimize the intracluster distance. At the same time, the aim is to maximize the separation between groups, to make the distance between the different clusters as large as possible. The result is a structure in which the clusters are homogeneous internally and heterogeneous externally.

Figure 2.1 provides a visual illustration of clustering applied to a two-dimensional dataset comprising two variables: height (x-axis) and weight (y-axis). The graph on the left shows the dataset before the algorithm is applied, with all observations shown in black and no grouping. The graph on the right shows the result after clustering has been applied, with the algorithm identifying two clusters: Cluster 1 in red, consists of individuals with lower heights and weights, and Cluster 2 in blue, consists of taller and heavier individuals. The algorithm groups observations according to similar patterns.

Figure 2.1: Cluster assignment visual representation



2.1.1 Clustering method

As it is a popular method in many fields, a wide variety of clustering algorithms can be found in the literature. The two best-known methods are:

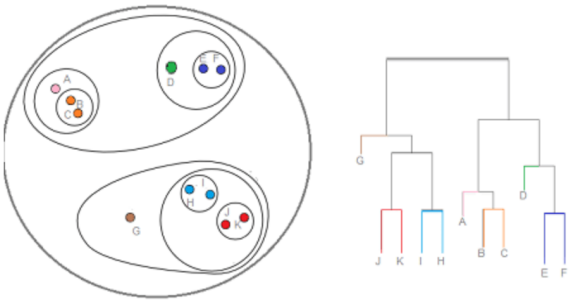
Hierarchical clustering The hierarchical *clustering* algorithm (Guess and Wilson, 2002) groups data by constructing a similarity tree, known as a *dendrogram* (Negri, 2022), as illustrated in

Figure 2.2. In this graph, the left-hand side shows a representation of the clusters, with individual points grouped within circles according to their proximity. On the right, the *dendrogram* shows the hierarchical equivalent of this grouping, where each branch represents a merger between groups or observations.

Initially, each observation forms a cluster on its own. Then, according to a similarity criterion, the closest groups are joined by branches. The vertical length of each branch in the *dendrogram* indicates the distance between the merged clusters: the longer the branch, the more different the groups are.

To obtain a final partition, the *dendrogram* is cut at a certain height. The number of branches remaining below the cut determines the number of final clusters.

Figure 2.2: Illustration of hierarchical clustering process and dendrogram output



Partitioning clustering The partitioning *clustering* algorithm (Swarndeep Saket and Pandya, 2016) divides a data set into a predefined number of clusters in a single pass, optimizing it according to a specific criterion to identify the most optimal partitioning of the data.

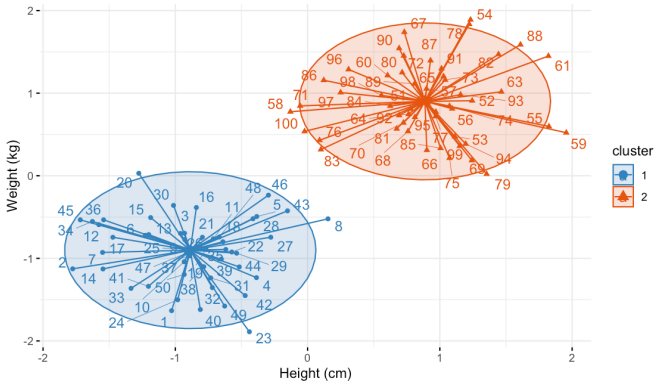
One of the best-known and most widely used methods within this approach is the K-means algorithm (Jin and Han, 2017). Starting with a pre-established value of K, the algorithm divides the data into K clusters. Each observation belongs to the cluster whose centroid is closest to it, as illustrated in *Figure 2.3*. This example shows a case of two clusters, $K = 2$, in blue and orange, formed from height and weight variables, with arrows connecting each observation to its centroid. The aim is to make the observations within each cluster as similar as possible to each other, while ensuring they are as different as possible from the observations in other clusters.

The K number of clusters is not known apriori and must be determined based on the characteristics of the data. Once K has been established, the algorithm seeks to minimize the intracluster variance, i.e. the sum of the squared distances (d) between each observation and its corresponding centroid. Let c_i be the centroid of cluster i, $\{x_{ij}\}$ the set of observations classified into this cluster, and j the index of the individual $(j = 1, ..., n_i)$. The criterion to be minimized is given

by:

$$\sum_i \sum_j d(x_{ij}, c_i)^2 \tag{2.1}$$

Figure 2.3: K-means clustering applied to a two-cluster scenario



There are several criteria for selecting the optimal number of clusters, K. Two of the best-known and most widely used of these are:

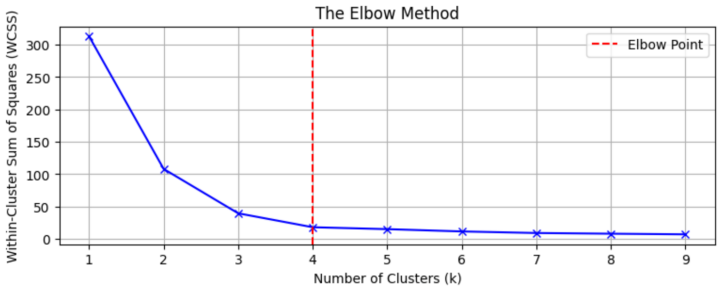
Elbow method (Humaira and Rasyidah, 2020): The elbow graph plots the within-cluster sum of squares (WCSS) values on the y-axis against the different values of K on the x-axis. The optimal K value is where the graph forms an 'elbow', indicating that increasing K no longer significantly reduces WCSS, as shown in *Figure 2.4*. In this example, the elbow appears at $K = 4$, indicating that a model with four clusters gives a good balance between complexity and intra-cluster compactness. Although this criterion is easy to implement, identifying the elbow can sometimes be subjective. The equation for WCSS is as follows:

$$\text{WCSS} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left\| x_{ij} - c_i \right\|^2 \tag{2.2}$$

Where:

- K: number of clusters
- n_i: number of data points in cluster i
- x_{ij}: j-th data point in cluster i
- c_i: centroid of cluster i

Figure 2.4: Elbow method for determining optimal number of clusters



Silhouette Index (Mamat et al., 2018): The Silhouette index can identify which objects are well placed within their cluster and which are somewhere in between clusters. The average Silhouette value is obtained for each K, and the value that maximizes this is chosen to indicate better separation between clusters. While this criterion provides a more objective measure, it is computationally more expensive. The equation for Silhouette Index is as follows:

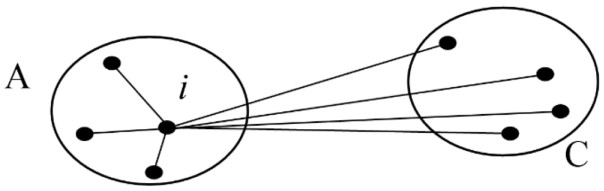
$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.3}$$

Where:

- $S(i)$ is the silhouette coefficient of the data point i
- $a(i)$ is the average distance between i and all the other data points in the cluster to which i belongs
- $b(i)$ is the average distance from i to all clusters to which i does not belong

Figure 2.5 shows how the Silhouette index is calculated for a given point i. It illustrates the average distance to points in the same cluster A ($a(i)$) and to points in the nearest neighboring cluster C ($b(i)$).

Figure 2.5: Illustration of Silhouette coefficient calculation



In the absence of clear criteria for determining the optimal number of clusters, the R package *NbClust* (Charrad et al., 2014) evaluates up to 30 different criteria and proposes the most appropriate number based on the majority criterion. However, this strategy is computationally expensive, particularly in a simulation study.

This study will apply the K-means clustering method. Although hierarchical clustering allows hierarchical structures to be visualized and does not require the number of clusters to be set in advance, the calculation and storage of distance matrices greatly increases the computational cost and time required as the size of the data grows. This makes it an nonviable choice.

The K-means method requires a criterion for deciding on the value of K. The Silhouette criterion will be used because it provides an objective and intuitive way of selecting K, without the need for additional assumptions about the shape of the data or subjective decisions such as the graphical visualization of the elbow method.

2.1.2 Cluster comparative metrics

This section presents the metrics used to compare the characteristics of the clusters obtained from real and synthetic data.

Percentage of equality of number of clusters In some cases, the algorithm may detect a different number of clusters in real data than in synthetic data. Therefore, this study is restricted to cases where both sets have the same number of clusters. This allows the synthesis quality analysis to be focused on and ensures that all metrics compare equivalent objects.

To avoid bias in the interpretation of the graphs or values obtained due to differences in the amount of data, a common number of necessary valid cases is set for the study scenarios. Next, the total number of cases that had to be generated to obtain this set number of real and synthetic data pairs with the same number of clusters is counted.

Based on this count, the percentage of equality in the number of clusters is calculated as the ratio between the fixed valid cases and the total number of cases generated.

Gini coefficient The Gini coefficient is a measure of inequality that quantifies the degree of imbalance in the distribution of a given resource. This study uses the Gini coefficient to determine the structure of the clusters by quantifying how balanced the cluster sizes are.

The coefficient takes values between 0 and 1. A value of 0 corresponds to perfect equality, meaning that all clusters contain the same number of observations. A value of 1 corresponds to perfect inequality, meaning that most observations are concentrated in one cluster, while the others are practically empty.

To compare the partition obtained with real data to the partition obtained with synthetic data, the absolute difference between their respective Gini coefficients will be used. A G value close to 0 indicates similar distribution of observations between the two clusterizations.

$$G = |G_R - G_S| \tag{2.4}$$

Where:

- G_R is the Gini coefficient corresponding to the real dataset.
- G_S is the Gini coefficient corresponding to the synthetic dataset.

Silhouette coefficient As explained, it is a measure that assesses the quality of a clustering partition by quantifying how well the observations within a cluster are assigned compared to those in other clusters.

One possible comparison metric is the absolute difference between the silhouette coefficients obtained in both clustering.

$$S = |S_R - S_S| \tag{2.5}$$

Where:

- S_R is the mean Silhouette index corresponding to the observations of the real dataset.
- S_S is the mean Silhouette index corresponding to the observations of the synthetic dataset.

A S value close to zero indicates similar quality of cohesion and separation between the two clusters, whereas a high value indicates differences in cluster definition between the two partitions.

Mean - Mean distance between centroids In each clustering, the centroid (the mean of the cluster values) of each cluster is determined. Then, the optimal correspondence between the centroids of the clusters from both clusterizations, with real and synthetic data, is found. Once the clusters have been matched, using the Hungarian algorithm (Mills-Tettey et al., 2007) (see section label switching), the Euclidean distance is calculated, and the mean of these distances is obtained and averaged.

To evaluate the structural similarity between clusters obtained from the original and synthetic data sets, the euclidean distance between the respective centroids is calculated. First, the centroid of each cluster is determined and then an optimal assignment is made between the centroids of the two sets, linking each real-data cluster to its closest synthetic counterpart. *Figure 2.6* shows a graphical representation of this situation. Once paired, the euclidean distance between each pair of centroids is calculated and the average of these distances is taken as an overall measure of similarity. The formula for the mean distance between centroids in the case

of having only two variables is as follows:

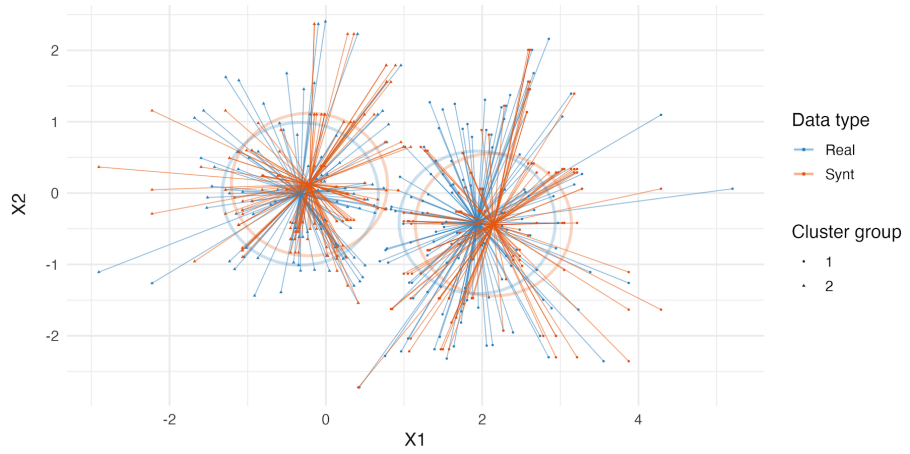
$$D(P_i, Q_i) = \frac{1}{K} \sum_{i=1}^{K} \sqrt{(p_{i1} - q_{i1})^2 + (p_{i2} - q_{i2})^2} \tag{2.6}$$

Where:

- K is the number of clusters.
- $P_i = (p_{i1}, p_{i2})$ is the centroid of cluster i in the real dataset.
- $Q_i = (q_{i1}, q_{i2})$ is the centroid of cluster in the synthetic dataset matched with cluster i of the real dat
- D_{centroid} is the average centroid distance between the two datasets.

A $D(P_i, Q_i)$ value close to zero suggests that the centers of the clusters are similar in both sets, indicating that the original structure has been well preserved.

Figure 2.6: Representation of centroid distances between real and synthetic data



Variance - Mean distance of variances To evaluate the similarity of variance in the clustering of real and synthetic data, one possible calculation is the mean variance distance, similar to the calculation used for mean centroid distance.

First, the variance of each cluster is calculated for both the real and synthetic data sets. Next, the optimal pairing between the clusters in the two sets is determined. Once this correspondence has been established, the absolute difference between the variances of each pair of clusters is calculated. Finally, the mean of the absolute differences is calculated.

$$D_{\text{var}} = \frac{1}{K} \sum_{i=1}^{K} \left| \sigma_{R,i}^2 - \sigma_{S,i}^2 \right| \tag{2.7}$$

Where:

- K is the number of clusters.
- $\sigma^2_{R,i}$ is the variance of cluster i in the real dataset.
- $\sigma^2_{S,i}$ is the variance of cluster in the synthetic dataset matched with cluster i of the real data..
- D_{var} is the average variance distance between the matched clusters.

A value D_{var} close to 0 indicates that the clusters in both sets have very similar variances and that the internal dispersion of the groups has therefore been correctly preserved.

2.1.3 Label switching

In clustering algorithms, a phenomenon known as *label switching* may occur. This happens when different runs of the same algorithm on the same dataset produce partitions with the same structure, but with different cluster labels. For instance, a cluster that was previously labeled 1 might be labeled 2 in a new run. This is because the numerical labels assigned to the clusters have no inherent meaning, they are simply identification labels assigned by the algorithm. This random change in numbering does not alter the structure of the groups, but it makes it difficult to compare results. This could be problematic in the present study, which aims to analyze the similarity between the clustering of real and synthetic data.

To ensure valid comparisons between the two sets, cases of label switching must first be identified and corrected. Otherwise, metrics such as distance or variance between centroids could produce biased results, creating false differences due to labeling rather than the real structure of the groups.

Figure 2.7 illustrates this problem, showing that before applying the label switching correction, the real and synthetic clusters appear to be swapped. In the real data, the blue cluster is associated with the orange cluster in the synthetic data, and vice versa. *Figure 2.8* shows the corrected labels of the corresponding clusters in each set.

Figure 2.7: Cluster assignment before label switching correction

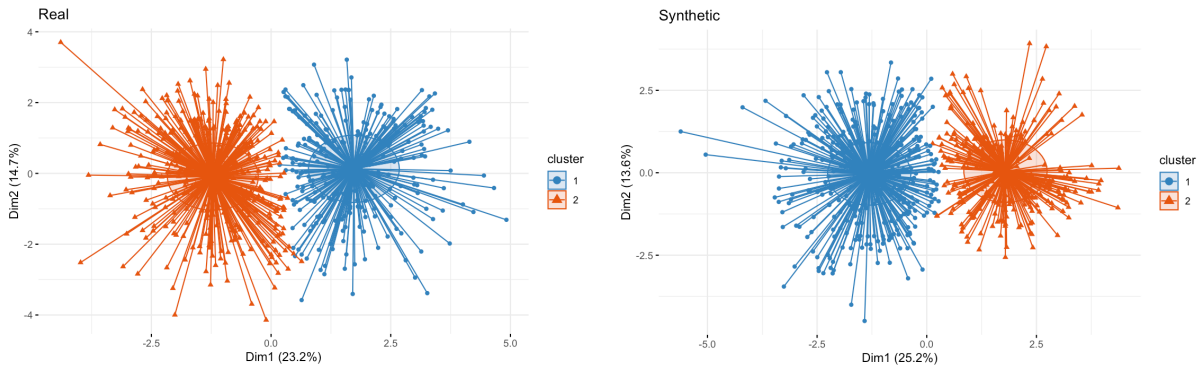
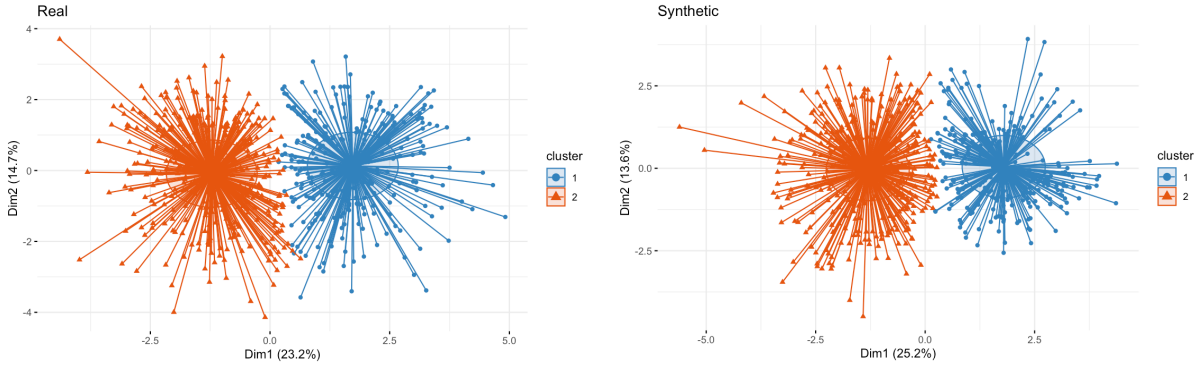


Figure 2.8: Cluster assignment after label switching correction



Label correction involves finding a correspondence between clusters in a way that maximizes the overlap of the assigned points within each cluster or minimizes the sum of the distances of the centroids, among other methods. Once the correspondence has been found, the clusters should be relabeled in the synthetic data so that they align with the clusters in the real dataset.

This correction is carried out using a function that employs the Hungarian algorithm. This algorithm solves the optimal assignment problem by pairing n agents with n tasks in a way that minimizes the total cost of the assignment. In this study, the algorithm is used to pair n real clusters with n synthetic ones and the cost to be minimized corresponds to the Euclidean distance between the centroids. The Euclidean distance between the centroids of the real and synthetic data is calculated for all possible combinations as shown in the example of the *Table 2.1*. Using the distance matrix, the Hungarian algorithm searches for the one-to-one assignment that minimizes the sum of these distances. For the case in *Table 2.1*, the possible one-to-one assignments are:

- Real 1 $\rightarrow$ Synthetic 1 i Real 2 $\rightarrow$ Synthetic 2, with cost $0.473 + 0.349 = 0.822$
- Real 1 $\rightarrow$ Synthetic 2 i Real 2 $\rightarrow$ Synthetic 2, with cost $3.031 + 3.001 = 6.032$

The algorithm detects the first assignment has a lower overall cost, i.e. the distance between the centroids is smaller. Then, a case of *label switching* is detected and the synthetic data clusters are relabeled to align with the real clusters.

Table 2.1: Distance between centroids

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

2.2 Synthetic data

Synthetic data (Raghunathan, 2021) is artificially generated data that does not come from real-world events, but is designed to resemble a genuine data set. It mimics the statistical properties, correlations and patterns of the original set, reproducing its structure and the interaction between its variables.

This process is possible because the synthetic data generation model learns the statistical characteristics and distribution of the real data, using them to create new records that are not linked to any real individual.

One of its main applications is the protection of privacy, as it enables data to be processed while maintaining the analytical value of the originals without exposing personal or sensitive information. Synthetic data also solves the problem of data scarcity, which is a challenge in the development of artificial intelligence systems as these require large volumes of information for efficient training.

2.2.1 Synthetic data generation method

Several approaches can be used to generate synthetic data in different contexts, depending on the characteristics of the data and the available resources. The appropriate methodology depends on the type of data and the objectives of the analysis. Three such approaches are presented below.

- **Based on statistical distribution** This approach involves generating synthetic data based on statistical knowledge of how real data is distributed, and a new set that exhibits the same statistical behavior or pattern is created.

  For instance, if a variable is known to follow a normal distribution, new data can be generated that also exhibits this characteristic. This method can be applied to other known distributions, such as exponential, gamma, among others in the case of continuous distributions.

  Provided the behavior of the data is well understood, this method can also be useful when real data is unavailable.

- **Based on an agent model** This approach involves creating a model that can learn from observed real-world behavior and use this information to generate new data. This is achieved by aligning real data with recognized statistical distributions, such as normal, exponential, chi-square distributions, among others. Once this adjustment has been made, the model can generate synthetic samples that follow the same statistical structure.

Alongside classical statistical methods, machine learning algorithms are employed to shape the distribution of the data. Techniques such as *CART* (Classification and Regression Trees) (James et al., 2013) and *CTree* (Nowok et al., 2016) can capture relationships between variables without making assumptions about their distribution. However, techniques such as decision trees can suffer from overfitting, whereby they adapt too much to the real data and lose predictive power.

- **Based on deep learning** This approach uses neural networks to reproduce the distribution of data. It is particularly useful for complex data, such as images, signals and unstructured data, as well as mixed sets containing both categorical and numerical variables.

  Well-known models in this approach include GANs (Generative Adversarial Networks) (Goodfellow et al., 2014) and VAEs (Variational Autoencoders) (Doersch, 2016). While they offer great flexibility and the ability to capture complex relationships between variables, they require large amounts of data and incur high computational costs.

2.2.2 Resemblance metric

The SPECKS statistic is a similarity measure designed to assess the quality of a synthetic dataset generated from a real dataset using a propensity score approach.

Based on the idea that synthetic data should not differ from real data, if the two are merged into a single set and a new indicator variable is added that takes the value 0 for real observations and the value 1 for synthetic observations, it is expected that training a classification model in which the indicator variable acts as the response variable and the others predict the probability that each observation is synthetic will enable the probability of each observation being synthetic to be predicted *Table 2.2*.

Table 2.2: Dataset on the left containing two variables and the true label (real or synthetic). Dataset on the right containing the same two variables and the propensity scores.

V1	V2	Label
160.7181	50.814 13	0
159.8003	49.533 74	0
160.2217	48.344 18	0
...	...	...
180.2575	64.494 33	1
180.4691	49.858 38	1
181.1581	65.646 33	1

V1	V2	Propensity score
160.7181	50.814 13	0.2995173
159.8003	49.533 74	0.2950036
160.2217	48.344 18	0.3176499
...	...	...
180.2575	64.494 33	0.5864138
180.4691	49.858 38	0.7624997
181.1581	65.646 33	0.5929186

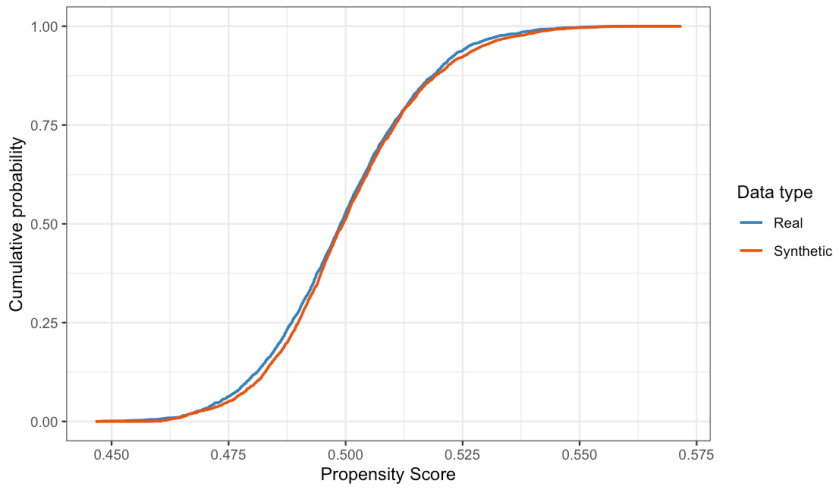
The predicted probability for each observation is known as the propensity score. If the synthetic data generation is correct, the propensity scores of the real and synthetic observations are expected to behave similarly. In other words, the synthetic data will not exhibit differential

patterns compared to the real data, and the model will not be able to distinguish which set each observation comes from. SPECKS therefore serves as a tool to assess the similarity between the cumulative distribution functions of the propensity scores of each set, by calculating the maximum absolute difference between them.

The Kolmogorov–Smirnov (KS) test calculates the maximum absolute difference between the cumulative distribution functions of two samples. This maximum value, known as the KS statistic, measures the discrepancy between the two distributions. The same concept is applied to propensity scores in the case of the SPECKS statistic *Figure 2.9*. Therefore, the SPECKS statistic follows a KS distribution and takes values between 0 and 1, as cumulative distribution functions take values within this range.

$$\text{SPECKS} = \sup_{\hat{p}_i} \left| \hat{F}^o(\hat{p}_i) - \hat{F}^s(\hat{p}_i) \right| \overset{H_0}{\sim} KS(n_o, n_s) \tag{2.8}$$

Figure 2.9: Cumulative distribution of Propensity Scores



The closer the distributions of the propensity scores are to each other, the closer the maximum difference will be to 0. A value close to 1, on the other hand, would indicate a large discrepancy between the two distributions. Once this maximum distance has been calculated, it is important to establish whether it is large enough to consider the two distributions different, taking the number of samples into account.

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

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

The theoretical distribution of SPECKS corresponds to the Kolmogorov–Smirnov distribution, enabling exact solutions to be obtained for the test.

The *stats* package implements the *ks.test* function, which calculates the KS statistic to compare two distributions and returns both the value of the statistic and the associated p-value. By applying this function to the propensity scores of the real and synthetic sets, the value of the statistic and the p-value can be obtained, thus facilitating the analysis of whether there is a significant difference between the two sets.

2.3 Tools for assessing SPECKS-variable relationship

Once the values of the clustering and similarity metrics described above have been collected, the next step is to analyze possible influencing factors and study the relationship between variation in the SPECKS value, similarity in clustering characteristics and the target. This section details the tools used to detect the relationship between SPECKS and the different metrics of clustering similarity.

Box-plot A box-plot is a graphical representation that illustrates the distribution of a numerical variable. It shows the mean, quartiles, extreme values and outliers of a dataset, which is useful for comparing distributions between different groups or variables.

Box-plots will be suitable because all clustering metrics are continuous numerical variables.

Box-plots will be constructed for each factor analysis and for each clustering metric. A visual comparison of the boxes will indicate whether differential patterns in the behavior of the clustering metrics exist at different levels of the same factor.

Correlation plot Correlation plots show the linear relationship between two variables.

In this study, they will be used as a visualisation tool to detect how SPECKS varies with respect to each of the clustering metrics. SPECKS is assigned to the X-axis and the different metrics are represented on the Y-axis. This allows patterns in the relationship between SPECKS and the metrics to be detected visually, revealing the extent to which SPECKS is related to clustering similarity.

Linear regression Linear regression is a statistical tool that enables the relationship between two numerical variables to be modeled and quantified.

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2.9}$$

The dependent variable Y will be the comparison metric and the independent variable X will be the SPECKS value. This enables us to analyze how the metrics value varies in relation to the

SPECKS values. This line is plotted on the correlation graph.

In addition to this visual approach, which can be subjective, numerical metrics will also be applied.

Pearson correlation coefficient Index that measures the linear relationship between two quantitative variables, independent of their scale.

$$R = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{2.10}$$

Where:

- σ_{XY} is the covariance of (X, Y)
- σ_X, σ_Y are the standard deviations of the marginal distributions

Slope of the regression line The slope shows how much the clustering metric changes on average when SPECKS increases by one unit. A positive slope suggests that higher SPECKS values are associated with higher-quality metrics. If β_1 is very close to zero and not statistically significant, this implies that there is no clear linear relationship between SPECKS and the metric under consideration.

Confidence interval of the slope of the regression line The confidence interval for β_1 serves to assess whether the slope obtained in the previous section is statistically significant different from 0. If the interval does not contain zero, then β_1 is significant, meaning that changes in SPECKS are associated with changes in cluster similarity. Conversely, if zero lies within the interval, the data cannot rule out the possibility that the slope is actually zero, meaning that it cannot be concluded that a significant relationship exists.

2.4 Simulation

The objective of the simulation set out in this study is based on two main lines of analysis.

Firstly, it aims to identify the factors that determine the similarity between the clusters obtained from real and synthetic data. Secondly, to detect factors in the data set that influence the replication of the original cluster structure for better or worse.

Secondly, based on the previous results, the study aims to define simulation scenarios to observe SPECKS' behavior and analyze its predictive capacity with regard to the quality of the clustering replication.

This section provides a detailed description of the steps involved in the simulation process, including the generation of the original and synthetic data.

2.4.1 Original data

A data generation function has been developed to allow different combinations of the factors considered to be simulated. All variables are generated from a multivariate normal distribution.

The possible influencing factors on the generation of the original analyzed data are as follows:

- Sample size: Our hypothesis was that the larger the original sample, the more accurately the synthetic data generation algorithm will capture the underlying structure of the data.
- Variable number: A larger number of variables are expected to provide more useful information for identifying patterns and relationships within the data.
- Degree of correlation: The impact of uncorrelated data and moderate correlation between variables on the clustering results will be studied.
- Number of clusters: Evaluate how the number of groups influences the quality and consistency of the clustering results.
- Separation between clusters: The distance between cluster centroids is controlled (*Figure 2.10*). The greater the separation, the easier it would be for the algorithms to correctly identify the groups.

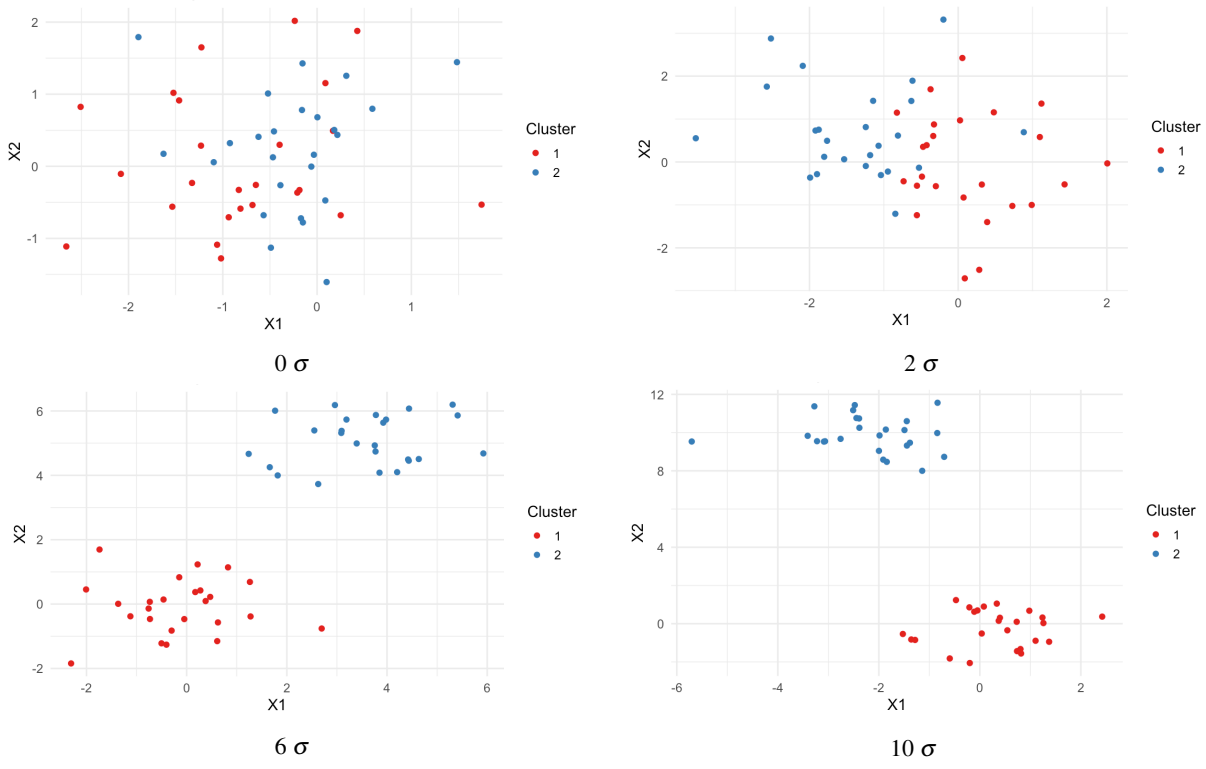
Data are simulated for all possible combinations of factor levels. Once the data have been generated, a uni-variate analysis is then applied to each cluster metric according to the considered factors, using box-plots.

The total number of possible combinations results in 432 different scenarios, which incurs a high computational cost. During the simulation and analysis process, some irrelevant levels or factors will be discarded. The factors analyzed and the expected results are as follows (*Table 2.3*):

Table 2.3: Summary of the experimental design factors, the specific values tested for each, and the expected influence they may have on the resemblance between real and synthetic data.

Factor	**Tested Values**	**Expected Effect**
Sample size (N)	50, 250, 1000	The larger the sample size, the better the resemblance between real and synthetic data.
Number of variables (p)	2, 5, 10	More variables may provide richer information for identifying patterns.
Degree of correlation (ρ)	0, 0.4	Correlated data may improve clustering consistency.
Number of clusters (k)	2, 3, 4	The more clusters, the challenger for the algorithms to replicate structure
Cluster separation (σ)	0, 2, 6, 10	The greater the separation, the easier to detect and replicate clusters.
Synthetic data method	CART, Normal	Different methods may affect how well the structure is captured.

Figure 2.10: Representation of cluster separation based on the distance between their respective centroids, expressed in terms of standard deviations σ.



A function has been developed to generate data that allow scenarios with specific characteristics to be simulated. This function controls various parameters, such as sample size, number of clusters, number of variables, correlation between variables, and separation between groups.

Before generating the data, a function is established to ensure a predefined minimum separation between clusters. This distance is defined in terms of standard deviations, as shown in *Figure 2.10*. The degree of overlap is determined by how many deviations there are between the centers of the clusters. The first centroid is placed at the origin, while the remaining centroid locations are determined to guarantee the specified minimum separation. This control enables scenarios with different levels of overlap between groups to be simulated, thus facilitating the analysis of situations involving clearly separated clusters as well as those that are more difficult to differentiate.

A correlation matrix is then constructed to define the relationships between the variables within each cluster. This matrix has a constant value, ρ, outside the diagonal, indicating homogeneous correlation between all pairs of variables. Based on this matrix and assuming a standard deviation of 1 for all variables, the covariance matrix is derived. This is then used to generate the multivariate data.

Once the centroids and the covariance matrix have been defined, observations for each cluster are generated using a multivariate normal distribution. Each cluster has its own centroid as its mean, but they all share the same correlation structure between variables. The number of observations per cluster is distributed equally among the total number of samples.

2.4.2 Synthetic data

The influential factor analyzed in the synthetic data generation process is the generation method used. This study considers both a parametric and a non-parametric approaches.

In this study, the original data is known to come from a normal distribution. Given this, it is reasonable to expect the parametric *Norm* generation method to provide the best results, as it corresponds to the actual structure of the data. However, the *Norm* method is only applicable when it can be assumed that the data follows a normal distribution — a condition that does not always hold true in real contexts.

For this reason, the study also analyses a non-parametric generation method, which does not require any assumptions about the distribution of the original data. The aim is to compare the results obtained using both methods in order to assess the extent to which the non-parametric method can replicate the data structure and produce similar clustering metrics regardless of its distribution.

Among the existing non-parametric methods, two of the best-known non-parametric generation

methods are *CTree* and *CART*. These methods, which are based on decision trees, allow the relationship between variables to be estimated without having to assume a prior distribution. This offers greater freedom for future studies that are not based on normal distributions.

In terms of their differences, both *CART* and *CTree* are classification tree methods, but they diverge in the partitioning criteria they use to construct the tree. *CART* is biased because it selects divisions based on impurity reduction using the Gini index, which favors variables with many unique values. However, *CTree* uses statistical contrasts based on p values to determine the best partition, and it allows this bias to be controlled.

Nevertheless, *CART* is characterized as a fast, simple, easily interpretable model. In this study, placing greater emphasis on variables with higher variability can be advantageous, as the objective is to preserve the original patterns. It is for this reason that the *CART* method is selected for implementation in synthetic data generation processes.

There are currently well-established and implemented methods for generating synthetic data. In this study, the R package *Synthpop* is used, as it is a flexible tool for generating synthetic data from original data using different methods.

2.5 Software

The *Table 2.4* lists the software packages and libraries used to carry out the experiments and analyses presented in this study. For the full list of packages used, please refer to the appendix.

Table 2.4: R packages used and their purpose in the study

Package	**Main purpose**
`synthpop`	Synthetic data generation.
`mvtnorm`	Generation of multivariate normal distributions.
`factoextra`	Clustering analysis calculation and visualization.
`clue`	Correction of label switching using the Hungarian algorithm.
`stats`	Kolmogorov-Smirnov test `ks.test()`.
`dgof`	Kolmogorov-Smirnov test for distribution comparison.
`DescTools`	Calculation of the Gini index.
`cluster`	Calculation of the Silhouette index.
`NbClust`	Determination of the optimal number of clusters.
`ggplot2`	Creation of plots and data visualizations.
`tidyverse`	Data manipulation and transformation.
`dplyr`	Efficient manipulation of tabular data (filtering, grouping, summarizing).
`scales`	Formatting scales in plots (percentages, labels, etc.).

3 Results

3.1 Influence of Factors on the Clustering Quality of Synthetic Data

The primary objective of this study is to assess the factors that influence similarity in clustering structure results derived from either original or synthetic data. This analysis will determine which characteristics should be considered when selecting a real dataset for clustering analysis purposes involving synthetic data.

3.1.1 Matching number of clusters analysis

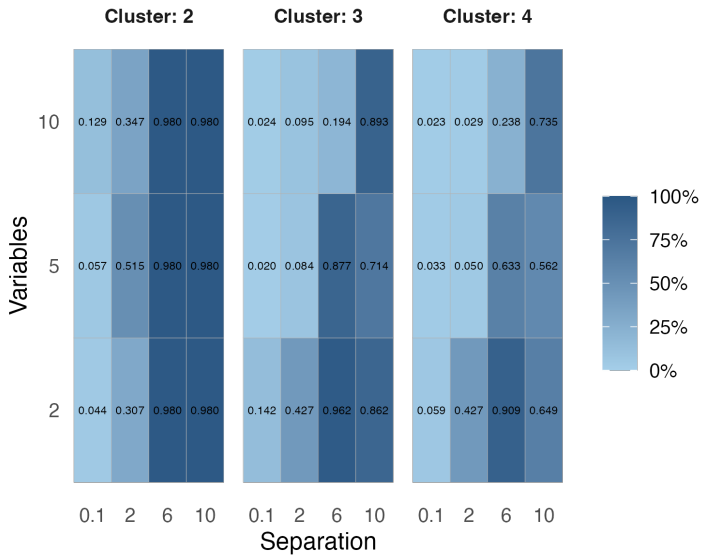
The analysis in this study is limited to cases where the number of clusters detected in the real and synthetic data match. Before simulation, the required number of observations to perform the analysis is set, and synthetic data sets are generated accordingly. Only those cases in which the number of detected clusters in the synthetic data matches that of the original data are retained. The number of iterations required to obtain the desired number of matching samples is recorded during the process. The ratio of valid cases to total iterations acts as a measure of structural consistency, indicating the extent to which synthetic data reproduces the cluster structure of the original data. The aim is to identify the conditions under which this match is most likely to occur.

Figure 3.1 illustrates the proportion of cases in which the number of clusters detected in the synthetic data matches the actual number of clusters for various combinations of factors, such as cluster separation, the number of variables, and the number of clusters. The separation between clusters has a clear influence on the variation in this proportion. Where the groups tend to overlap, as with small separations, the proportion of agreement between the real and synthetic cluster numbers is low. Conversely, as the separation increases — as in cases involving six or ten standard deviations, for example — the percentage of agreement rises considerably. This pattern is visually reflected by an intensification of blue tone from left to right. This suggests that greater separation between groups allows the original data to be distributed more accurately into groups by the synthetic data generation algorithm. This result was expected.

Regarding the number of clusters, it should be noted that the more groups present in the data, the harder it becomes to achieve a match. For example, with two well-separated clusters, the proportion of matches is relatively high, whereas this proportion decreases in scenarios with four clusters. The number of variables does not significantly affect the proportion of matches on its own. However, it interacts noticeably with the separation factor. In data with few variables, achieving a high proportion of matches does not require a very high separation. In contrast, as the number of variables increases, greater separation is required to achieve a high degree of

similarity between real and synthetic data.

Figure 3.1: Proportion of matching cluster number between real and synthetic data, across different separations, number of variables, and number of clusters



This study aims to analyze the quality of synthetic data in clustering analysis. For the purposes of this study, cases have been limited to those with the same number of clusters. This restriction enables the use of metrics to compare equivalent objects. The results presented in the subsequent sections are based solely on cases with matching numbers of clusters.

3.1.2 Factors behavior according to clustering metrics in different scenarios

To reduce the computational cost of the study, it is important to note that the total number of combinations increases exponentially with each additional factor. Given this, this study examines whether the choice of synthetic data generation method (*Norm* vs. *CART*) significantly affects the behavior of clustering metrics. This analysis helps determine whether both generation methods must be used throughout the study, or if one can be excluded to optimize computational resources.

Figure 3.2 illustrates the comparison of this two generation methods with respect to the values of the four proposed clustering metrics: the difference between the real and synthetic Gini coefficients (*Diff.gini*), the difference in silhouette coefficients in both datasets (*Diff.sil*), the mean distance between centroids (*Mean.distance*) and the mean difference in cluster variance (*Mean.var*). The results show that, for three of the four analyzed metrics, *Diff.gini*, *Mean.distance* and *Mean.var*, there are no significant differences between the methods. The distributions are practically identical in both the quartiles and the dispersion. This suggests that the two methods are comparable in their ability to preserve the relative size of the clusters, the

distance between centroids, and the internal variability of the groups.

The *Diff.sil* metric does show a notable difference between the two methods. The *CART* method shows values closer to zero and less dispersion, which could suggest greater stability and consistency in the compactness of the generated clusters. However, the distribution is slightly shifted towards negative values and does not include zero within the confidence interval. This could suggest a slight tendency to overestimate the cohesion of synthetic clusters compared to real ones. By contrast, the *Norm* method exhibits significantly greater variability in this metric and includes the value 0 within the confidence interval. This suggests that despite its dispersion, it has the capacity to generate cluster structures that resemble real ones in certain instances. However, this greater variability may also indicate that synthetic data tends to produce less compact clusters. In other words, while it reproduces the general shape of the original clusters, it generates more dispersed groupings.

Figure 3.2: Comparison of Clustering Quality Metrics by Synthetic Data Generation Method (*CART* & *Norm*)

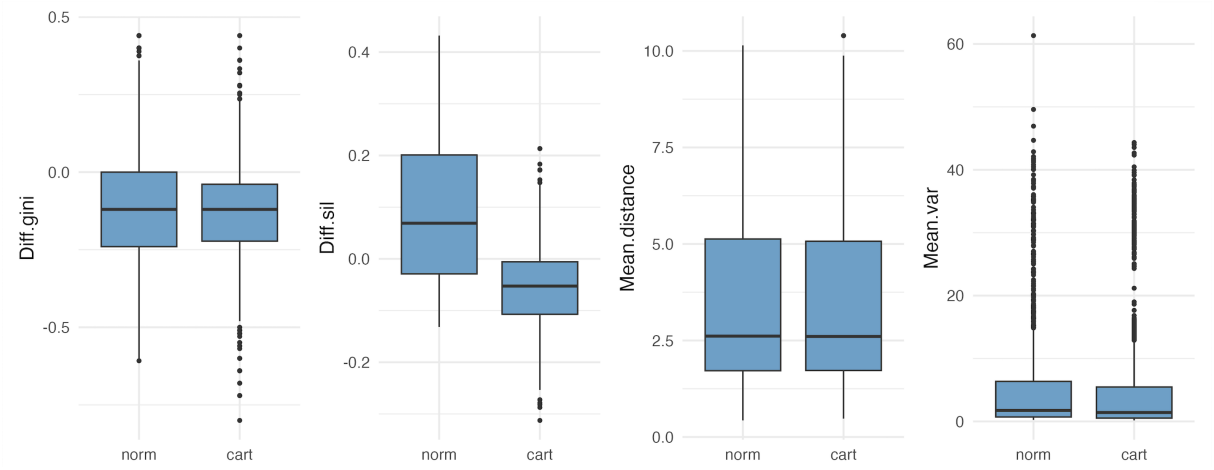


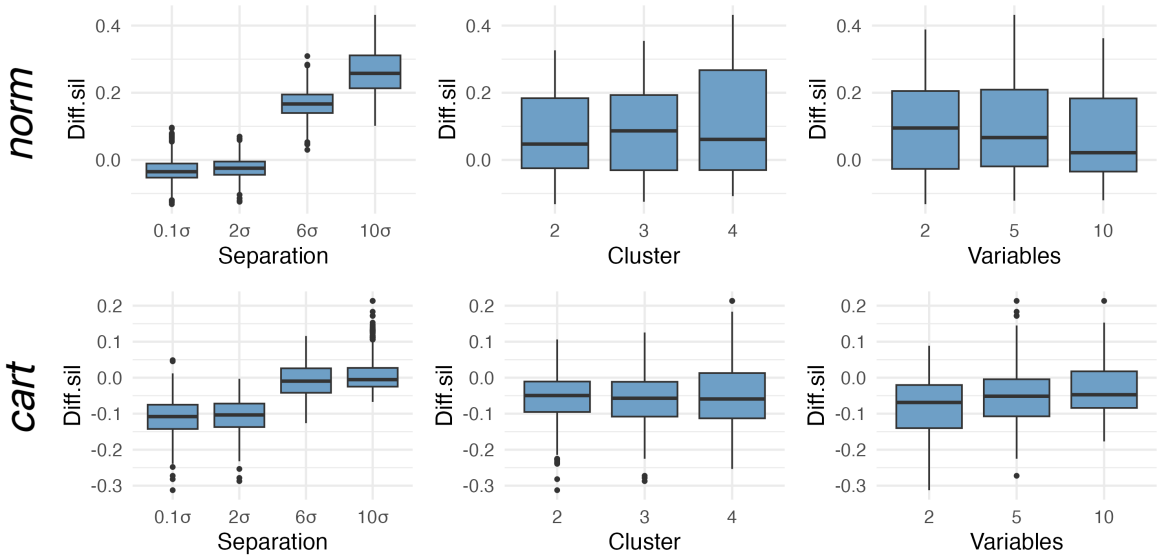
Figure 3.3 continues the analysis of the *Diff.sil* metric, showing how it evolves based on three factors: cluster separation, cluster number and variable number. This allows the difference between the two generation methods to be detected more clearly.

Regarding the separation factor, both methods show a noticeable evolution, albeit with different behaviors. In the case of the *Norm* method, an increase in separation between clusters results in a progressive increase in *Diff.sil* values, which is counterintuitive. This behavior suggests that despite the clusters being more clearly defined in the real data, the synthetic data generated can result in more dispersed and less compact groupings. In contrast, the *CART* method shows a reduction in *Diff.sil* values as separation increases, resulting in less dispersion. This pattern indicates a greater ability to preserve the structure of the original clusters when they are moderately differentiated. However, when the separation is very high, i.e. 10 standard deviations, a significant increase in outliers is observed, suggesting that the method is less stable in situations

of extreme separation.

No differential pattern is observed for either method with regard to the number of clusters. Both Norm and Cart maintain relatively stable *Diff.Sil* values depending on the number of groups, suggesting that this factor does not significantly affect the quality of the generated groups. Similarly, no notable differences were identified in the number of variables. The *Diff.sil* values demonstrate a consistent trend across the various levels of dimensionality considered (two, five and ten variables), indicating that an increase in complexity does not significantly impact the comparability between real and synthetic clusters.

Figure 3.3: Effect of Cluster Separation, Number of Clusters, and Variables on *Diff.sil* by Generation Method



Despite some minor differences observed in certain metrics, no significant differences were detected between the results obtained using the two generation methods. In terms of preserving the structure of the clusters, both methods exhibited similar behavior in most of the analyzed cases. The execution times of each method are compared below to assess their relative computational efficiency.

Table 3.1 shows the computational times for the two generation methods for one iteration running. *User* time is the time that R uses to execute the code; *System* time is the time that the operating system uses to manage resources; and *Elapsed* time is the actual time that the whole process takes.

The *CART* method is clearly more efficient, with a total execution time of 255.2 seconds compared to 325.2 seconds for the *Norm* method in average. This difference is also reflected in the user and system times, indicating that the *CART* method requires fewer computational resources to generate the same volume of synthetic data.

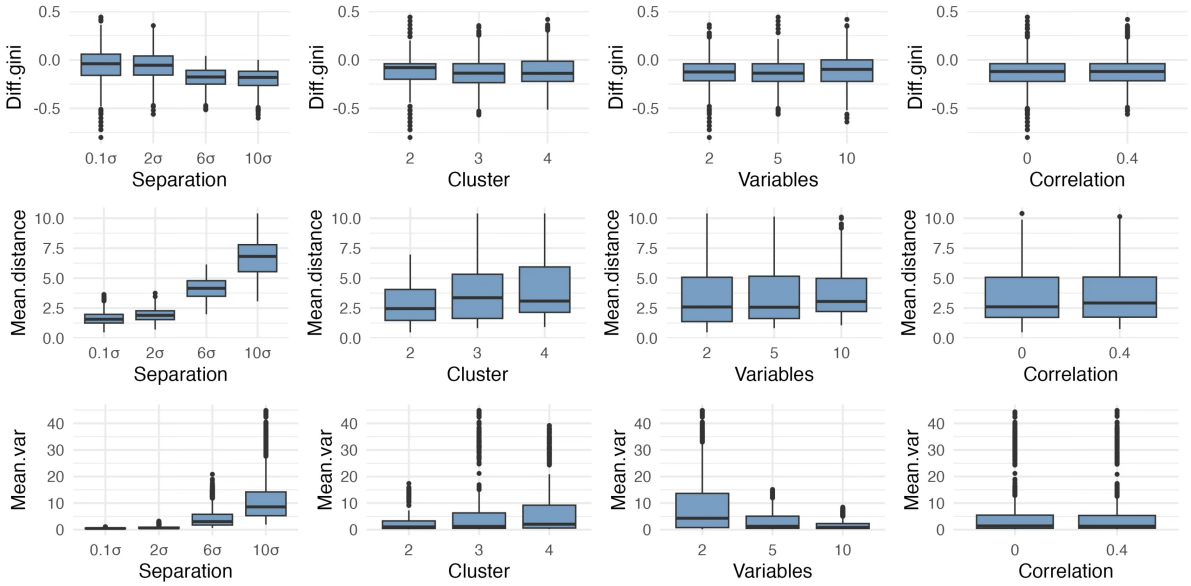
Table 3.1: Execution time for 1 iteration

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

As no significant differences were found between the results of the two methods and the execution time of the Cart method is clearly shorter than that of the Norm method, the following analyses will be based solely on the Cart method results, with the aim of optimizing computing resources.

Figure 3.4 illustrates the impact of the aforementioned factors on the various clustering metrics. Each row of the graph represents a metric, and each column corresponds to a factor. The metrics analyzed are the difference in Gini impurity (Diff.gini), the mean distance between cluster centroids (Mean.distance), and the mean variance difference between clusters (Mean.var). The factors analyzed are the level of separation between clusters, the number of clusters, and the number of variables in the database. The sample size for the combinations shown in this graph is 50 observations.

Figure 3.4: Distribution of clustering metrics (*Diff.gini*, *Mean.distance*, *Mean.var*) according to four factors: cluster separation, number of clusters, number of variables, and variable correlation.



Diff.gini The first row of the graph illustrates the behavior of *Diff.gini*, if the clusters have similar structures, they are expected to have similar Gini values and therefore their difference will be close to 0. It is noteworthy that most of the values are negative. This is because the real data are generated with balanced clusters. Therefore, if the clustering obtained from the

synthetic data shows a slight deviation with one or two changes to the observation assignment, this results in a negative Gini difference.

Focusing on the *Separation* factor, when the distance between centroids is small, as in cases of 0.1 or two standard deviations, the value of *Diff.gini* tends to be closer to zero, and sometimes exceeds it. However, this pattern no longer holds when the clusters are more distant, for example at 6 or 10 standard deviations. Although the original data is designed to generate balanced clusters, the clustering algorithm may not always be able to distinguish groups perfectly in situations of overlapping clusters, which explains the presence of positive values in *Diff.gini*. Conversely, when the separation is very large, the algorithm correctly detects the balanced clusters in the original data; however, this clear pattern may not be reproduced in the synthetic data. In this case, small deviations in the assignment of some observations can generate lower Gini variation, resulting in a smaller difference and lower variability.

No clear differential patterns are observed for the *Cluster* and *Variables* factors between the different levels, indicating that these two factors do not significantly impact the variation of the *Diff.gini*.

Mean.distance Regarding the *Separation* factor, if the centroids are similar in both datasets, the mean distance is expected to be close to zero. It is observed that, for data with closer clusters, the average centroid distance tends to be smaller when buying datasets where the clusters are more clearly differentiated.

As the number of clusters increases, the variability of the mean distance also increases slightly, as observed in the *Cluster* factor.

As for the *Variables* factor, no differential patterns are identified between the different levels. However, a slight decrease in variability of the metric is observed as the number of variables increases, particularly in the presence of outliers, i.e. in some cases, there may be clusters with a significant difference in centroid position between real and synthetic data.

Mean.var If the clusters are similarly variable in both datasets, the mean variance difference is expected to be close to 0. A clear increasing trend is detected for the separation factor. When the clusters are in close proximity to each other, as in cases involving a separation of 0.1σ or 2σ , the difference in variability between the real and synthetic data is small. However, as the distance increases, as in cases involving a separation of 6σ or 10σ , the *Mean.var* value and its dispersion also increase. Greater separation amplifies the internal variability of the clusters between the two types of data.

Regarding the *Cluster* factor, the greater the number of clusters, the greater the variability of *Mean.var*. With more clusters, there are more deviations in variability. In contrast, for the

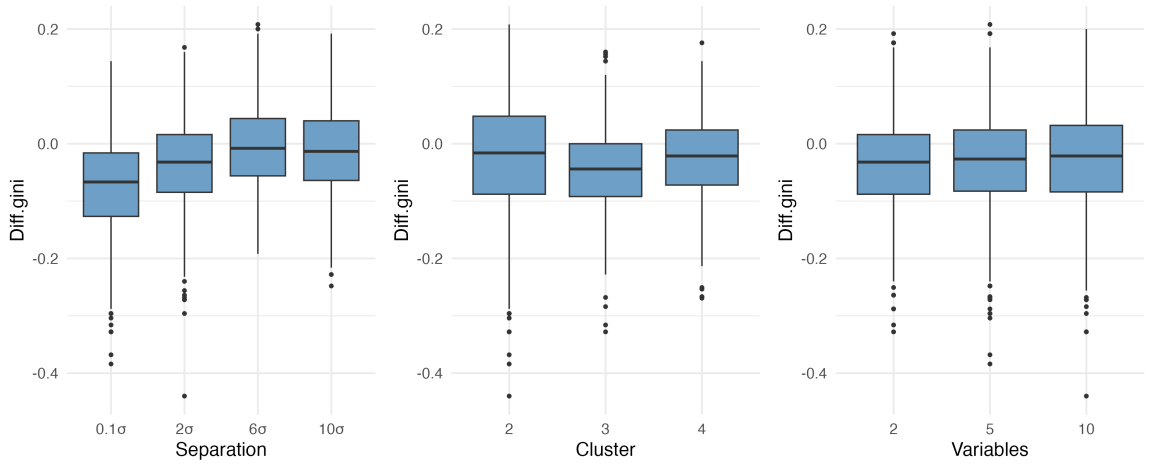
Variables factor, a decreasing trend in the metric value is observed. When the data have larger dimensions, synthetic generation is more likely to reproduce the variability structure of the original clusters accurately.

Looking at the last column of graphs, corresponding to the analysis of the two-level correlation factor, $\rho = 0$ and $\rho = 0.4$, it does not seem to have a relevant effect on any of the clustering metrics studied. In all cases, the distribution of metric values for both correlation rings are practically identical. The degree of correlation between variables does not have an impact on the similarity of the clusters generated from the synthetic data.

3.1.3 Gini behaviour for unbalanced cluster

In the previous case, where the data had been created using balanced clusters, the values of *Diff.gini* tended to be below 0 for the reason abovementioned. This raises the question of whether the synthetic data can correctly preserve the structure of the original clusters. *Figure 3.5* analyses the case of unbalanced clusters to determine whether this remains true.

Figure 3.5: *Diff.gini* values across different levels of separation, number of clusters, and number of variables in unbalanced cluster scenarios. The x-axis represents the varying factor; the y-axis shows the difference in Gini index between real and synthetic data



In this unbalanced scenario, the *Diff.gini* metric tends to be closer to 0 than in the previous case. This approximation to 0 is most clearly observed in cases involving well-separated clusters, where the mean *Diff.gini* is centered around 0, and the synthetic generation method maintains the distribution of cluster observations.

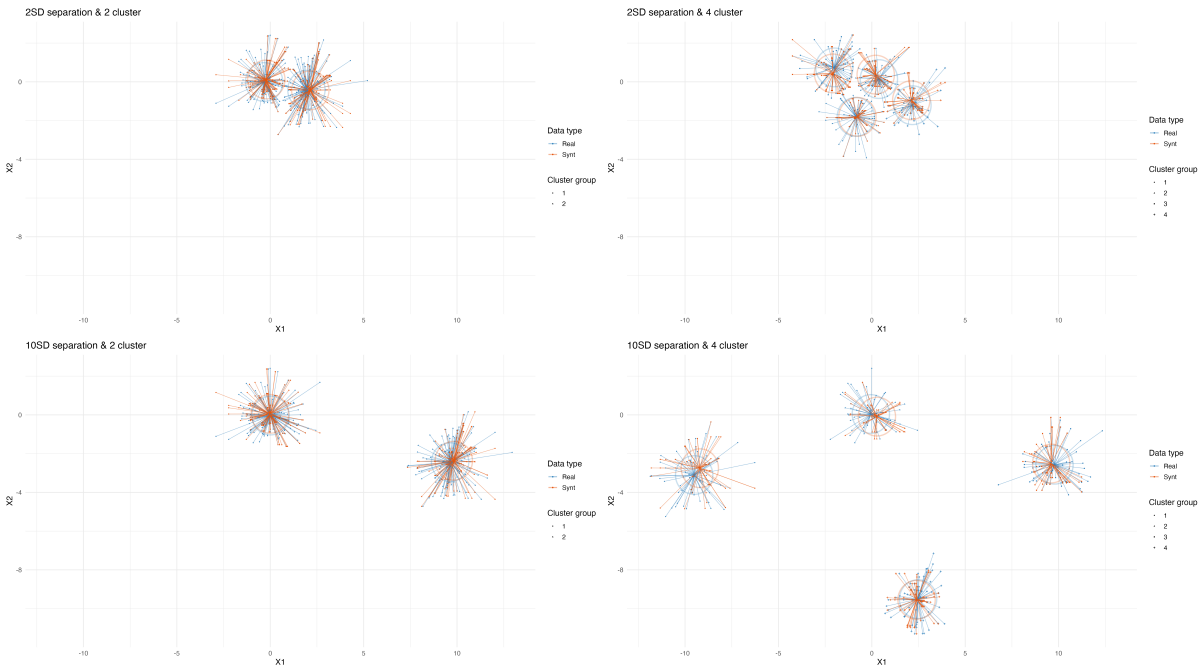
3.1.4 Cluster centroid behaviour

In analysing the *Mean.distance* and *Mean.var* metrics, it had suggested that an increase in the number of clusters in the data would result in an increase in both the mean distance between centroids and the difference in variability within clusters. *Figure 3.6* provides a graphical representation of these results, showing the real and synthetic clusters superimposed in the same space. This allows us to analyze how the clusters behave in terms of group separation and shape, as a function of the number of clusters and the distance between centroids.

Results obtained from the *Mean.distance* metric showed a progressive increase in average centroid distance as group separation and the number of clusters increased. *Figure 3.6* shows that, as the separation between clusters increases, the clusters appear further apart visually. At a separation of 2σ, the real and synthetic clusters overlap, while at a separation of 10σ, the clusters appear more isolated. Similarly, increasing the number of clusters shows this pattern of increasing distance between cluster centroids. This confirms the results obtained in the *Mean.distance* metric.

This graphical representation has been rescaled to a uniform level across the graphs to facilitate direct comparison between the different scenarios. A non-rescaled version of the figure has also been generated, where the absolute magnitudes of separation and dispersion can be seen more clearly. This figure is included in the appendix of the work in order to provide a more complete view of the geometric behavior of the clusters.

Figure 3.6: Comparison of Real and Synthetic Cluster Centroids



3.2 SPECKS: Using a Resemblance Metric as a Proxy for Clustering Utility

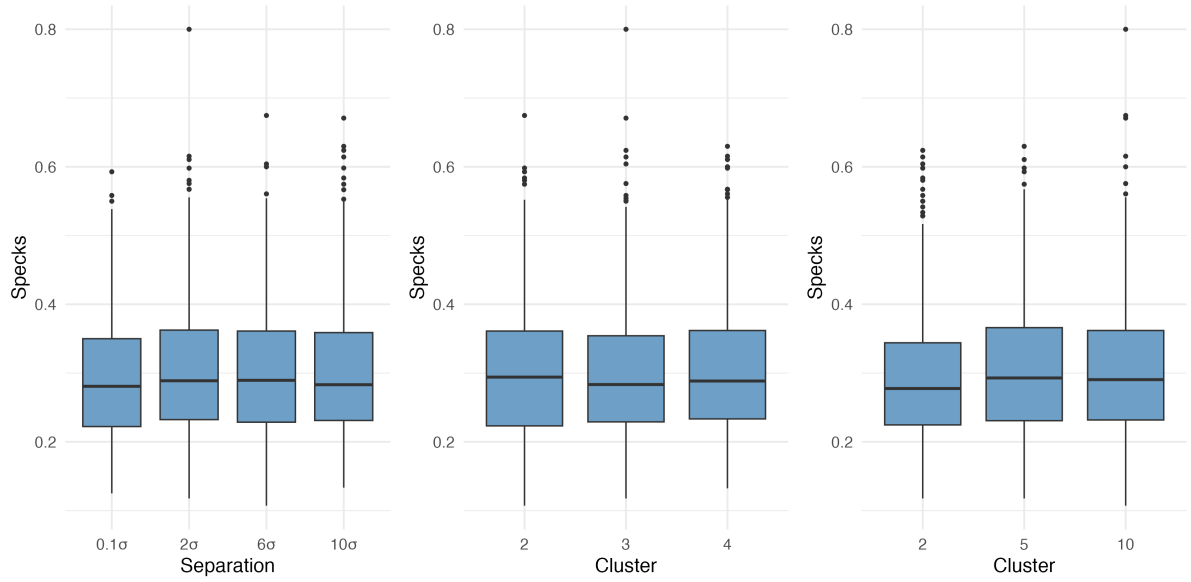
The second objective of this study is to assess the suitability of the SPECKS assembly metric for clustering analysis. The aim is to analyze its ability to capture the quality of synthetic data relative to real data in terms of preserving cluster structure. The interest is to determine whether SPECKS can function as a reliable metric for assessing the extent to which synthetic data maintains the necessary structure for clustering analysis. This section will therefore study SPECKS' behavior as a reliability indicator and examine its relationship with clustering metrics.

3.2.1 Factors

Figure 3.7 illustrates the distribution of SPECKS in relation to three factors: cluster spacing, the number of clusters, and the number of variables. These are the same factors that were examined above. This analysis enables us to examine how SPECKS behave at different levels of each factor and whether they respond coherently to elements that impact the structure of the data, according to the clustering metrics.

The factor analysis in the previous section provided information on which factor levels influence the similarity of the generated clusters to the real ones. Based on this, SPECKS metric values would expect to reflect these structural differences.

Figure 3.7: SPECKS values across different levels of separation, number of clusters, and number of variables. The x-axis shows the varying factor; the y-axis represents the SPECKS metric.



It was observed that the mean difference between the cluster centroids for the *Separation* factor increases as the distance between them grows. If SPECKS can capture this change, its values

are expected to vary according to the level of separation. However, the graph shows that the SPECKS value remains constant at all levels of separation.

For the *Cluster* factor, metrics such as *Mean.distance* and *Mean.var* demonstrate increased variability as the number of clusters increases. Nevertheless, SPECKS maintains an equal distribution across two, three and four clusters. It does not respond adequately to the structural change of the clusters.

The *Variables* factor, the metric *Mean.var*, showed that, in the real data, increasing dimensionality led to greater similarity in terms of variability among the clusters in the synthetic data. The stability of SPECKS across all levels of variables indicates that it does not reflect this change.

The relationship between SPECKS and the various clustering metrics is analyzed using scatter plots and regression lines.

All possible combinations of the studied factors were considered, resulting in a total of 72 scenarios. A linear regression is fitted for each combination, analyzing the direction and statistical significance of the obtained slope.

Figures 3.8 - 3.11, depict a selection of the 72 graphs. For each metric, an attempt has been made to select a case for each of the following cases: one with a significantly positive slope, one with a significantly negative slope, and another with a non-significant slope. This finding indicates a lack of consistency in the results and suggests that the observed relationship between the variables may be attributable to chance rather than a systematic effect. Some graphs show a statistically significant trend, but in most cases the slopes of the regressions are not statistically significant and include the value 0 within the 95% confidence interval. Only a very small number of combinations show a significant slope at the 5% level, and these cases are rare and do not follow any consistent or repeatable pattern across scenarios. This suggests that the observed relationships may be due to randomness rather than a genuine relationship between SPECKS and cluster quality.

Figure 3.8: SPECKS values (x-axis) across and the difference in Gini index (y-axis) under three different cluster matching scenarios.

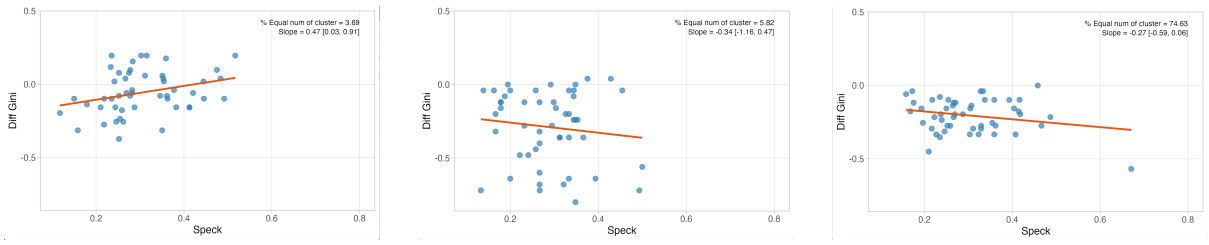


Figure 3.9: SPECKS values (x-axis) across and the difference in Silhouette index (y-axis).

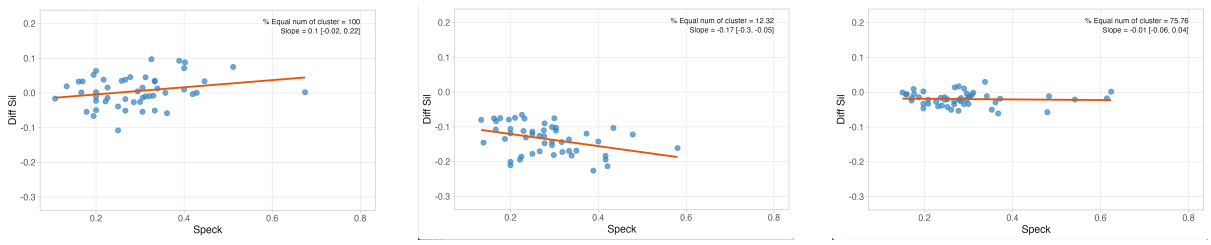


Figure 3.10: SPECKS values (x-axis) across and the difference in clusters mean distance (y-axis).

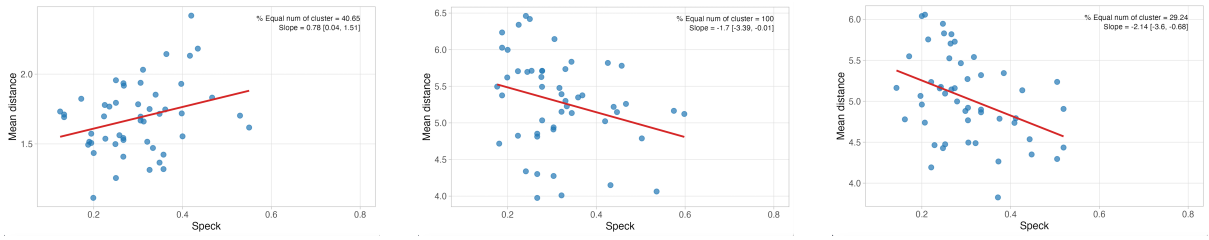
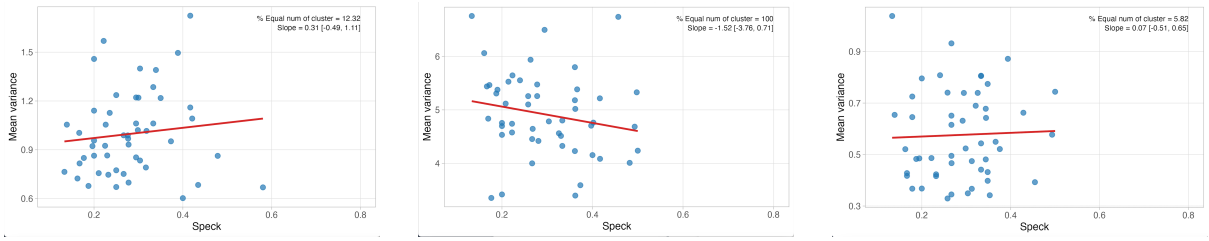


Figure 3.11: SPECKS values (x-axis) across and the difference in clusters mean variance (y-axis).



The value of SPECKS remains stable across all levels and factors, which raises concerns about its discriminative power. Its lack of variation in response to changes in cluster structure suggests that Specks is not sufficiently sensitive to serve as a reliable indicator of structural quality in the context of synthetic data clustering analysis.

3.2.2 Scenarios for further analysis

Based on the information obtained from the previous analysis, two scenarios have been established for further analysis. Firstly, an ideal scenario is defined in which real data are generated with factor values that, according to previous results, favor greater similarity in cluster structure between real and synthetic data. The second scenario is more problematic, characterized by a combination of factors that make it difficult to preserve this structure.

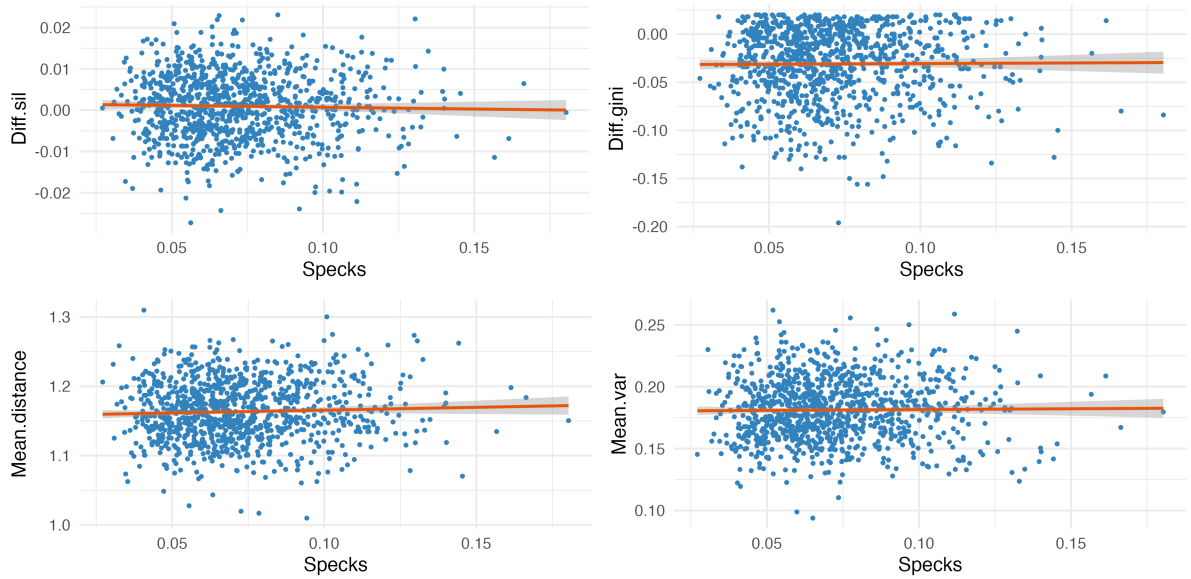
A database with 1000 observations and 8 variables was set for both the ideal and problematic cases, with the aim of providing sufficient data for the *cart* generation method to adequately

capture the characteristics and internal structure of the original data.

Ideal case The generated dataset contains two clusters separated by two standard deviations, i.e. with some overlap between groups. This scenario was selected because it combines elements that favor preserving the cluster structure, as indicated by the factor analysis results: moderate separation between groups and a small number of clusters. It is expected that the clusters of the generated synthetic data will maintain a similar structure to that of the real data.

Figure 3.12 shows the correlation results between the SPECKS and the four cluster comparison metrics between real and synthetic data in the ideal scenario. The Silhouette and Gini differences are both close to 0, indicating that the quality of clustering in both sets is similar. Similarly, the differences in mean distances and variances between clusters are stable and not very high. Regarding the relationship with SPECKS, all values are homogeneous, and the fit lines have a very low slope and no correlation pattern is evident. There are no significant correlations between SPECKS variation and the clustering metrics, which reinforces the idea that SPECKS does not evaluate the quality of synthetic generation in terms of reproducing the structure of the original data.

Figure 3.12: Relationship between SPECKS and structural difference metrics in a scenario with two moderately separated clusters (2σ), chosen for its stability and expected structure preservation.



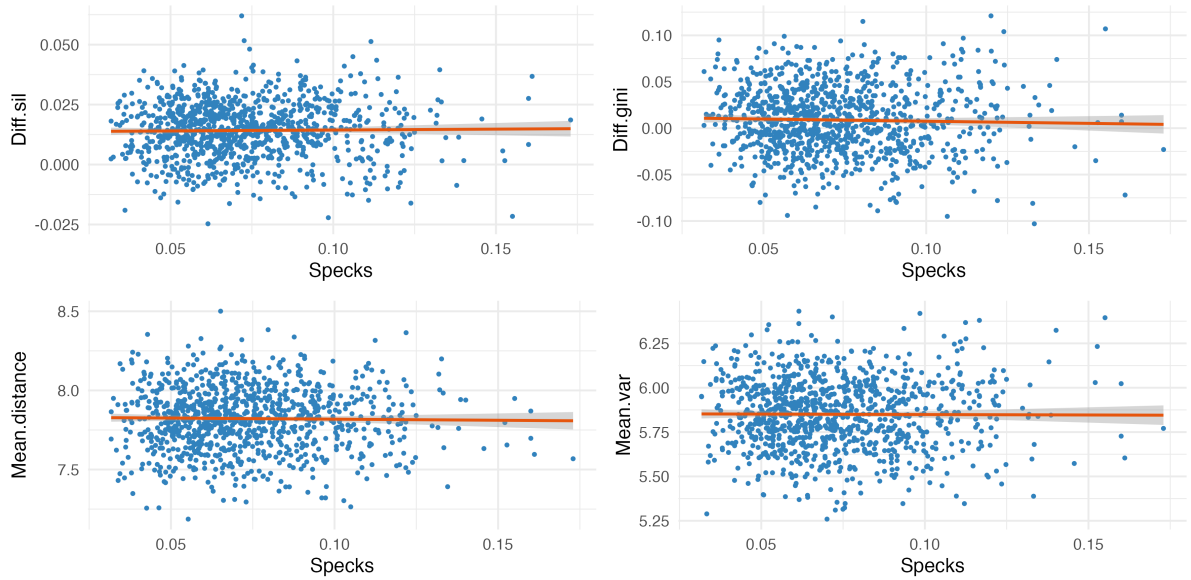
Problematic case The generated dataset contains four clusters, with a separation of ten standard deviations between centroids — i.e. the groups are differentiated. This scenario has been selected to analyze a situation in which the real data were generated under conditions that, according to the previous factor analysis, make it difficult to preserve the cluster structure in the

synthetic generation process, i.e. a clear separation between groups and a greater number of clusters. It is expected that the clusters in the generated synthetic data will not maintain the structure of the real data as well as in the ideal scenario described above.

Unlike in the ideal case, *Figure 3.13* shows that the comparative metrics between the real and synthetic data present values that are further away from 0, indicating a lower similarity in the structure of the clusters. However, SPECKS' scale of variation is maintained and its range of values is the same as in the ideal case. In this scenario, SPECKS also shows no relationship with respect to structural differences. There is no significant association between the SPECKS value and the variations of the metrics in any of the four graphs. The fit lines are flat with no relevant trends.

This result reinforces the conclusion from the ideal scenario: the SPECKS metric shows no sensitivity to changes in the structural quality of the synthetic data. Despite the poorer replication of cluster characteristics in the problematic case, SPECKS maintains the same range of value variation, showing almost constant values in situations where the other metrics vary. Therefore, SPECKS is not a suitable metric for assessing the quality of synthetic data when the goal is to preserve cluster structure.

Figure 3.13: Relationship between SPECKS and structural difference metrics in a scenario with high separation (10σ) and four clusters, chosen for conditions known to hinder structure preservation in synthetic data.



4 Discussion

The study focuses on analyzing the ability of synthetic data to reproduce the cluster structure observed in the original data set. The main objective is to evaluate whether the Specks assembly metric can be used to measure the quality of this structural replication.

To this end, a simulation has been designed in which data is generated under various conditions with different levels of the following factors: the number of clusters, the separation between groups, the correlation between variables, and the method of synthetic generation. The analysis was completed by studying the scenarios defined based on the results obtained in the factor analysis.

Key findings

First, certain factors have been identified to have a significant impact on the similarity between the clusters obtained from real and synthetic data. The most influential factor is the separation between clusters. As this increases, metrics such as the average distance between centroids and the average variance difference reveal greater discrepancies between the two datasets. One possible explanation for this is that when clusters are far apart, there is more space around them. This greater freedom can result in the original structure being reproduced less accurately. Making it more difficult to maintain the same configuration of clusters when they are clearly differentiated. It could be said that synthetic data generation reproduces the structure of the original data less accurately when the data groups are widely separated. Apart from that, it has also been found that a larger number of clusters hinders the ability of synthetic methods to correctly replicate the original structure. With more partitions, the resulting groups are smaller and more susceptible to internal variations, which can lead to greater discrepancies between real and synthetic data. The number of variables and the degree of correlation do not significantly impact the analyzed metrics.

Secondly, with regard to the ability of the Specks metric to indicate how well synthetic data preserves the structure of the original clusters, the results show that this metric is not affected by the quality of replication of cluster characteristics. The range of Specks values remains virtually constant regardless of changes in the analyzed factors, showing no significant relationship with the different metrics used to compare clustering quality. Specks values obtained in scenarios with good cluster replication, as defined by the clustering metrics, do not differ from those obtained in cases where replication is less effective. Therefore, Specks does not reflect differences in cluster structure between synthetic and real data. Its stable behavior means it is ineffective at detecting whether the characteristics of the original data clusters have been maintained in the analysis of synthetic data. Therefore, it can be concluded that Specks is not a suitable metric for assessing the structural quality of synthetic data in clustering analysis.

When generating synthetic data with the aim of preserving the structural patterns of clusters, it is relevant to use an original database containing a large number of samples and variables, and with a simple cluster structure comprising a small number of groups that are moderately separated from one another. This will lead to a better replication of the clusters' characteristics in the synthetic data.

Using Specks as a similarity indicator may be insufficient to ensure that the cluster structure has been maintained, as it does not adequately reflect the differences between real and synthetic data in this context.

Comparison with previous work

Few studies have examined the ability to preserve clusters in the context of synthetic data generation. A recent paper published in the *MDPI Electronics* journal (Petricioli et al., 2025) proposes a methodology for generating synthetic data that maintains the structure of the original clusters. This approach involves incorporating structural information from the clusters, such as quartiles and correlation matrices, with the aim of preserving the shape and internal distribution of the groups. The authors demonstrate that, without taking this structural information into account, clusters may disappear or become distorted, despite the synthetic data appearing to replicate well at the global level.

This study takes a similar approach, evaluating the extent to which synthetic data generation can reproduce the cluster structure of the original data. While the *MDPI Electronics* paper proposes a method to ensure this preservation, however, this study focuses on analyzing different scenarios to determine which factors in the original data influence the quality of structural replication in the synthetic data. Both papers agree that overall similarity between datasets is not sufficient to ensure good preservation of the internal structure.

Limitations

Despite the results, this study has several limitations that could lead to future research lines.

In terms of selecting the optimal number of clusters, the study used the Silhouette index, a well-known and widely used method, but there are many other criteria in the literature, such as the Elbow method, Calinski-Harabasz Index, Dunn Index, among others. There is no clear determination as to which criterion is best in general terms. An interesting alternative would have been to use the *NbClust* function in R, which calculates the number K using 30 different criteria and selects the most frequent value. This would allow for the optimal number of clusters to be estimated more accurately. However, this was not possible due to the high computational cost of simulating a large number of scenarios.

With regard to the sample size factor, three levels were considered: 50, 250 and 1000 observa-

tions. It was not possible to include scenarios with larger samples, such as 10000 observations. This would have allowed the behavior of the methods to be analyzed in contexts involving larger data volumes. This limitation is also mainly due to the high computational cost of generating and analyzing large data sets, since even for scenarios with 1000 observations, running the simulation proved challenging.

The study focused solely on Specks as a metric for measuring similarity between real and synthetic data, in terms of the similarity of the obtained clusters. While this approach enabled an initial evaluation, other assembly metrics, e.g. propensity mean square error (Snoke et al., 2018), that might better capture cluster structure preservation more effectively and detect internal structural differences between groups were not analyzed. Future work could incorporate and compare other metrics to assess whether synthetic data preserves the structure of the original clusters.

Generalization

The results of this study should be interpreted within the specific framework of the simulation and cannot be extrapolated to other contexts.

All the original data used for synthetic generation came from a multivariate normal distribution. Therefore, the effects observed in the different analyses are only valid in this context and cannot be generalized to cases where real data follow different types of distribution.

The analysis focused exclusively on cases where the same number of clusters was detected in both the synthetic and real datasets. While this approach ensured a direct comparison between the two datasets, it excluded scenarios in which this match did not occur. The impact of this exclusion on the quality of the replicated clustering was not analyzed.

The study focused solely on the context of cluster analysis, and all the comparison metrics used were designed to evaluate group structure preservation. Other types of statistical analysis, such as ANOVA, regression, etc., where the structural requirements of synthetic data may be different, were not considered. Therefore, the results presented cannot be considered representative for other analytical purposes.

Future research directions

Future studies could explore alternative metrics for assessing the relationship between metric values and the structural similarities in the clustering of real and synthetic data. These metrics should be more sensitive to cluster characteristics and structures. It would also be interesting to extend the analysis to data with distributions other than the normal distribution and to other statistical approaches, such as regression analysis. Greater computational resources would enable larger sample sizes to be evaluated, leading to more robust conclusions.

Conclusions

The work was centered on evaluating the ability of synthetic data to preserve the cluster structure of the original data, as well as assessing the usefulness of the Specks assembly metric for measuring this similarity. The results showed that the separation between groups and the number of clusters were determining factors in the replication quality. However, Specks has not been shown to be sensitive to differences in these clustering metrics, meaning it is not suitable for assessing the quality of structural replication in clustering analyses.

References

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.

Charrad, M., Ghazzali, N., Boiteau, V., and Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36.

Doersch, C. (2016). Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guess, M. J. and Wilson, S. B. (2002). Introduction to hierarchical clustering. *Journal of clinical neurophysiology*, 19(2):144–151.

Humaira, H. and Rasyidah, R. (2020). Determining the appropiate cluster number using elbow method for k-means algorithm. In *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*, pages 1–8.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jin, X. and Han, J. (2017). K-means clustering. In *Encyclopedia of machine learning and data mining*, pages 695–697. Springer.

Mamat, A. R., Mohamed, F. S., Mohamed, M. A., Rawi, N. M., and Awang, M. I. (2018). Silhouette index for determining optimal k-means clustering on images in different color models. *Int. J. Eng. Technol*, 7(2):105–109.

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*.

Negri, R. G. (2022). Dendrogram. In *Encyclopedia of Mathematical Geosciences*, pages 1–3. Springer.

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

OpenAI (2024). Chatgpt: Language model by openai.

Petricioli, L., Humski, L., and Vranić, M. (2025). Preserving clusters in synthetic data sets based on correlations and distributions. *Electronics*, 14(11):2230.

Polak, P. and Anshari, M. (2024). Exploring the multifaceted impacts of artificial intelligence on public organizations, business, and society. *Humanities and Social Sciences Communications*, 11(1):1–3.

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.

Swarndeep Saket, J. and Pandya, S. (2016). An overview of partitioning algorithms in clustering techniques. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 5(6):1943–1946.

5 Appendix

5.1 Supplementary Figures

Figure 5.1: Comparison of Cluster Centroids Between Real and Synthetic Data Across Varying Cluster Counts and Separation Levels

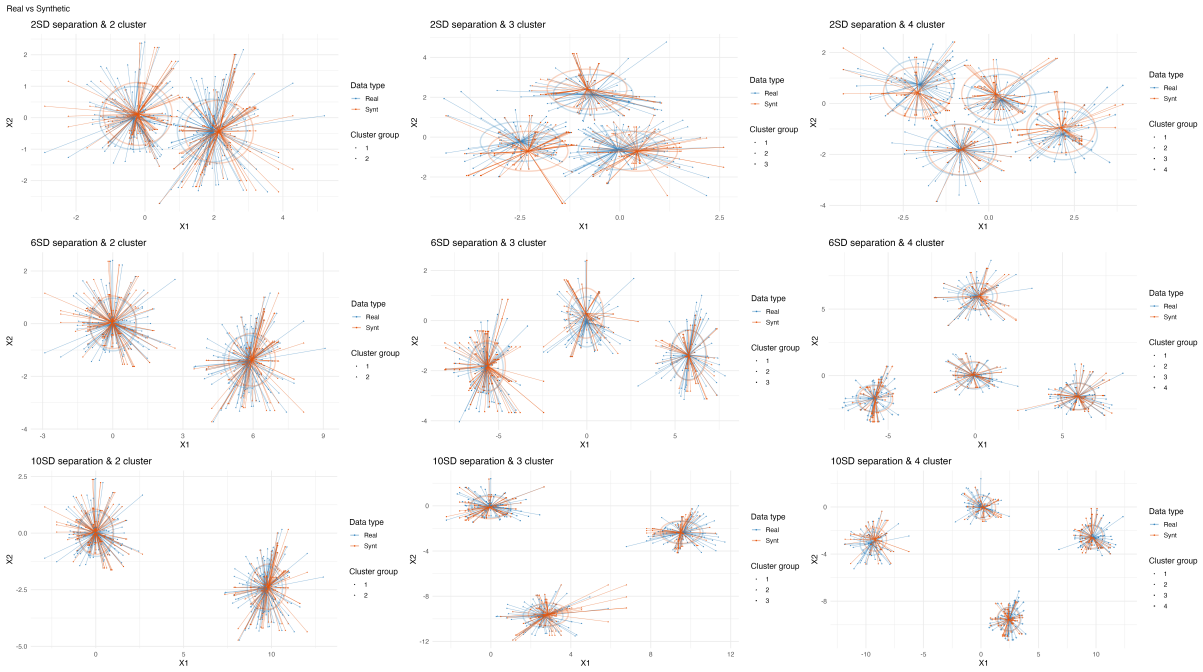


Table 5.1: R packages used and their purpose in the study

Package	Main purpose
`synthpop`	Synthetic data generation.
`mvtnorm`	Generation of multivariate normal distributions.
`MixSim`	Simulation of clusters with specified overlap.
`clusterGeneration`	Creation of clustered multivariate data.
`factoextra`	Clustering analysis calculation and visualization.
`clue`	Correction of label switching using the Hungarian algorithm.
`label.switching`	Handling label switching in Bayesian mixture models.
`stats`	Kolmogorov-Smirnov test `ks.test()`.
`dgof`	Alternative implementation of Kolmogorov-Smirnov test.
`DescTools`	Calculation of the Gini index.
`cluster`	Calculation of the Silhouette index.
`NbClust`	Determination of the optimal number of clusters.
`caret`	Classification and regression model training.
`rpart`	Construction of decision trees.
`car`	Diagnostic tools for regression analysis.
`ggplot2`	Creation of plots and data visualizations.
`patchwork`	Combining multiple `ggplot2` plots.
`paletteer`	Access to custom color palettes for plots.
`scatterplot3d`	3D scatterplots for visualizing clusters.
`scales`	Formatting scales in plots (percentages, labels, etc.).
`glue`	String interpolation for dynamic labels and messages.
`parallel`	Parallel computation to improve performance.
`tidyverse`	Data manipulation and transformation.
`dplyr`	Efficient manipulation of tabular data (filtering, grouping, summarizing).
`tidyr`	Data reshaping (pivoting, nesting, unnesting).
`tibble`	Enhanced printing and handling of data frames.

5.2 Code

The code used to carry out this work is available in the GitHub repository below: `https://github.com/XinnuoChen/Final-Work` .