



NOVA

IMS

Information
Management
School

MGI

Mestrado em Gestão de Informação

Master Program in Information Management

**OUTLIER DETECTION FOR IMPROVED
CLUSTERING**

Empirical research for unsupervised data mining

Jacob Hastrup Madsen (M2016087)

Dissertation for obtaining the Master's degree in Information
Management with a Specialization in Business Intelligence
and Knowledge Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

LOMBADA MEGI

2017

Outlier Detection for Improved Clustering
Empirical Research for Unsupervised Data Mining

Jacob Hastrup Madsen

MEGI

LOMBADA MGI

2017

Outlier Detection for Improved Clustering
Empirical Research for Unsupervised Data Mining

Jacob Hastrup Madsen

MGI



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

OUTLIER DETECTION FOR IMPROVED CLUSTERING

by

Jacob H. Madsen

Dissertation proposal presented as partial requirement for obtaining the Master's degree in Information Management with a Specialization in Business Intelligence and Knowledge Management

Advisor: *Prof. Dr. Ana Cristina Costa*

February 2018

ACKNOWLEDGEMENTS

To fellow students Jan Benedikt and Marcel Motta. Thanks for the good times of intensive studying during exam and project periods. A special thanks to Jan for always being supportive for whatever math or computational problem that would occur. One year in Lisbon is a year you will never forget.

To the Danish and European taxpayers, thank you for providing financial support and enabling me to go to Portugal and going on Erasmus semester in Germany.

ABSTRACT

Many clustering algorithms are sensitive to noise disturbing the results when trying to identify and characterize clusters in data. Due to the multidimensional nature of clustering, the discipline of outlier detection is a complex task as statistical approaches are not adequate.

In this research work, we contend that for clustering, outliers should be perceived as observations with deviating characteristics worsening the ratio of intra-cluster and inter-cluster distance. We present a research question that deals with improving clustering results specifically the for two clustering algorithms, k-means and hierarchical clustering, by the means of outlier detection. To improve clustering results, we identify and discuss the literature of outlier detection, and undertake on 11 algorithms and 2 statistical tests to the process of treating data prior to clustering. To evaluate the results of applied clustering, six evaluation metrics are applied, of which one metric is introduced in this study.

Using real world datasets, we demonstrate that outlier detection does improve clustering results with respect to clustering objectives, but only to an extent where data allows it. That is, if data contains ‘real’ clusters and actual outliers, proper use of outlier algorithms improves clustering significantly. Advantages and disadvantages for outlier algorithms, when dealing with different types of data, are discussed along with the different properties of evaluation metrics describing the fulfillment of clustering objectives.

Finally, it is demonstrated that the main challenge of improving clustering results for users, with regards to outlier detection, is the lack of tools to understand data structures prior to clustering. Future research is emphasized for tools such as dimension reduction, to help users avoid applying every tool in the toolbox.

KEYWORDS

Outlier Detection, Unsupervised Learning, Clustering, Data Mining,

INDEX

1. Introduction	1
1.1. Background	1
1.2. Problem	1
1.3. Research objectives.....	2
1.4. Structure of dissertation	3
2. Methodological framework	5
2.1. Literature review on outlier detection	5
2.2. Literature review on clustering.....	6
2.3. Application of outlier detection and clustering.....	6
3. Literature review	7
3.1. Methodology for identification of literature on outliers.....	8
3.2. Outlier detection approaches	9
3.2.1. Statistical Z-test.....	12
3.2.2. Robust statistics	13
3.2.3. Tukey's boxplot	13
3.2.4. Mahalanobis distance	14
3.2.5. Q-Q plot and Chi-square plot.....	16
3.2.6. Distance-based approaches.....	17
3.2.7. Density-based approaches.....	21
3.2.8. High-dimensional approaches	33
3.3. Clustering	38
3.3.1. Clustering objectives.....	38
3.3.2. Clustering evaluation	39
3.3.3. Cluster tendency	47
3.3.4. Clustering algorithms	49
4. Methodology for applying theory	52
4.1. Selection of datasets	52
4.2. Outlier detection and clustering analysis	54
4.2.1. Outlier detection settings	55
4.2.2. Clustering settings.....	56
4.3. Coherence for evaluation metrics.....	56

5. Results and discussion.....	58
5.1. Evaluation metrics.....	58
5.2. HeartDisease dataset	60
5.3. PimaIndians dataset.....	63
5.4. Frogs dataset.....	66
5.5. Ionosphere dataset	69
5.6. WineQuality dataset	72
5.7. Overall performance of outlier algorithms.....	75
5.7.1. Hierarchical clustering	75
5.7.2. K-means clustering.....	76
5.8. Understanding coherence and cluster results	77
5.9. Understanding removal percentage	82
5.10. Understanding advantages and disadvantages of outlier methods	82
6. Conclusion	85
7. Recommendations for future research.....	87
7.1. Connecting outlier detection and clustering	87
7.2. Connecting dimension reduction and clustering.....	87
8. References.....	88
9. Annex.....	92
9.1. Similarity matrices for datasets	92
9.2. Evaluation metrics for synthetic data treated with outlier methods	93
9.3. Comments on results for incoherent datasets	94
9.3.1. BreastCancerOriginal	95
9.3.2. ForestFires.....	98
9.3.3. BreastCancerDiagnostic	101
9.3.4. GesturePhase	104
9.4. Figures for incoherent results	107
9.4.1. BreastCancerDiagnostic	107
9.4.2. BreastCancerOriginal	113
9.4.3. ForestFires.....	119
9.4.4. GesturePhase	125
9.5. Figures for coherent results	131
9.5.1. PimaIndians.....	131

9.5.2. Ionosphere	137
9.5.3. HeartDisease	143
9.5.4. Frogs	149
9.5.5. WineQuality	155

ABBREVIATIONS

ABOD	Angle-based Outlier Detection
COF	Connectivity-based Outlier Factor
CRISP-DM	Cross-industry standard process for data mining
DB	Distance-based Outlier
FBOD	Feature Bagging Outlier Detection
INFLO	Influenced Outlierness
KNN	k-nearest neighbor
LOCI	Local Outlier Correlation Integral
LOF	Local Outlier Factor
LOOP	Local Outlier Probability
MAHA	Mahalanobis Distance
RNDM	Random
SOD	Subspace Outlier Degree
SSB	Sum of Squares Between
SST	Sum of Squares Total
SSW	Sum of Squares Within
t-SNE	t-distributed stochastic neighbor embedding

1. INTRODUCTION

1.1. BACKGROUND

The discipline of data mining has in recent years seen a significant increase in research contributions, making it a constant updating field of science. For data mining, unsupervised learning is the discipline of gaining insights in unlabeled data by the means of different tools, usually following a data mining process such as CRISP-DM (Shearer, 2000), SEMMA (SAS, 2006) or KDD (Fayyad, Piatetsky-shapiro, & Smyth, 1996). Among the tools available in unsupervised learning, clustering is one of the most frequent used. Following the above-mentioned data mining processes, data preparation/modification/preprocessing (mentioned respectively to the papers) is executed prior to modelling, as one cannot model better than data allows. For data preparation, the task of detecting and treating outliers is important, for clustering to return optimal results.

However, the task of detecting outliers is subject to discussion as it is a abstract question defining what an outlier is. Moreover, no universally accepted approach to outlier detection with regards to clustering results has been developed.

With the ever-expanding field of literature, several authors have over time developed different outlier detection methods, claiming to have developed the single solution or algorithm that produce the best outlier detection within a certain data domain (Filzmoser, Maronna, & Werner, 2006; H.-P. Kriegel, Kröger, Schubert, & Zimek, 2009; B. Tang & He, 2017). However, no research has been done connecting outlier detection to optimized clustering results.

1.2. PROBLEM

The no free lunch theorem states that no single algorithm produce the best result across all mathematical or computational problems. The same applies for clustering and outlier detection, where some clustering algorithms, such as DBSCAN (Ester, Kriegel, Xu, & Miinchen, 1996), have built-in exclusion of noisy objects, while other algorithms cannot deal with outliers.

Among the most used clustering algorithms, both k-means and hierarchical clustering lack properties to deal with outliers. Thus, as with any modelling problem, it is crucial to identify outliers for the clustering to return satisfying results, and based on allowance, either remove the outliers or reassign them after applied clustering. Returning to the no free lunch theorem, it is unknown how k-means and hierarchical

clustering excel in collaboration with certain outlier detection methods, and within what domains of data the two clustering algorithms have their respective advantage. This leads to following identified issues;

- Neither k-means nor hierarchical clustering is suitable for all kinds of datasets
- Neither k-means nor hierarchical clustering can detect and deal with outliers

Leading to the problem:

- The definition of an outlier is abstract. Depending on properties of data, it can be based on statistical tests, based on proximity, based on an algorithm, labeled or scored.

1.3. RESEARCH OBJECTIVES

The main objective of this thesis is to contribute to the literature of data mining and create insight in outlier detection with regards to improved clustering performance. To achieve this goal, the main research question to be answered in this thesis is:

- **How can outlier detection improve clustering results for k-means and hierarchical clustering?**

As implied in main research question, the thesis deals with two kinds of algorithms: outlier detection algorithms and clustering algorithms. The thesis is divided in to a theoretical part discussing tools identified in a literature review and a practical part applying the identified literature to data. To answer the main research question, a set of sub questions are formulated:

1. What is outlier detection and which tools does the literature provide?

Building foundational knowledge for the faced issue, this question serves to understand the nature of outlier detection and identify and understand the functionality of the outlier detection methods.

2. What is clustering, what are the objectives and how are clustering results evaluated?

To understand clustering and its objectives, a second literature review chapter will discuss the literature and describe the two clustering algorithms k-means and hierarchical clustering, later moving them on to the process of applied clustering. To understand how clustering objectives are quantified, clustering evaluation is also discussed.

3. How do outlier detection methods perform when comparing cluster results for data treated with different outlier detection methods?

This question is approached from a practical perspective and deals with the actual research results by applying theory to given data. The question of performance will be quantified from the perspective of clustering objectives identified in question two.

This thesis aims contribute to research by creating an understanding of advantages and disadvantages of outlier detection, for different clustering algorithms, for different datasets. If results return improved performance, it is expected that this thesis can help create an understanding of combinations of algorithms powerful for improved clustering performance.

1.4. STRUCTURE OF DISSERTATION

Chapter 1 outlines the motivational background for the thesis and the scope of reasearch. Any discussion raised in the thesis serves the overall purpose of answering the research question stated in this chapter.

Chapter 2 contains the general methodological framework of the thesis and describes the approach of answering the research question. It highlights how literature is discussed and literature is illustrated with practical cases. It also lists the general steps of applying theory of outlier detection and clustering to different datasets.

Chapter 3 presents the identified literature and is divided into three sections. The first section discusses the methodology for identification of literature. The second section deals with outlier detection and maps the contributions of literature until now. The literature is thoroughly described for the reader to understand technicalities. The third section deals with clustering and covers the topic by discussing objectives, tools and evaluation. Objectives of clustering are discussed with identified literature and serves to create a unified approach for later evaluation of clustering results. Evaluation of clustering results are described and discussed with identified evaluation metrics. A practical example is given for the reader to understand how evaluation metrics differ and to understand their respective advantages and disadvantages.

Chapter 4 describes the modelling methodology for applying theory to datasets. The first section discusses the selection of the datasets subject to applied theory. The importance of diversity in selected data is emphasised to answer the research question. The methodology for application of outlier detection and

clustering to selected datasets is detailed in the second section. Finally, coherence for evaluation metrics is discussed in the third section.

Chapter 5 presents and discusses the results. It is described how the large amount of produced metrics are handled for the reader to understand the overall result. Results for datasets are divided in two categories: positive results with improved clustering performance, and dubious results with questionable performance.

Chapter 6 presents the overall conclusion with respect to the research question. The conclusion highlights the outcome for the two groups of datasets and discusses major findings.

2. METHODOLOGICAL FRAMEWORK

2.1. LITERATURE REVIEW ON OUTLIER DETECTION

For the theoretical part dealing with outlier detection a literature review is conducted. The overall objective of the literature review is to identify secondary sources and map the current contributions to the discipline of outlier detection. The literature review strives to identify the most recognized algorithms to limit the number of algorithms applied to data, as testing every algorithm will create an unnecessary redundant long thesis. Furthermore, methodology for identification of literature is discussed at the beginning of the literature review chapter. To understand the nature of outlier detection and functionality of the outlier detection tools, the literature review contains discussions on a theoretical and practical level.

The theoretical discussion deals with the definition of an outlier and is rather philosophical. The literature is described from its early stages and discuss how the definition of an outlier has changed over time. The intuition of outlier detection, for identified approaches, is discussed for the reader to follow researchers' reasoning of their respective suggested technical approaches.

The practical discussion includes illustration examples of data developed by the respective authors in order to understand technicalities of the theory and to elaborate on advantages and disadvantages of the reviewed approaches. Focus is on four different properties: *applicability*, *method input parameters*, *method output* and *overall completeness*

- *Applicability* is the extent to which an outlier detection method can deal with datasets of different characteristics.
- *Method input parameter* is the subjective input parameters that algorithms use as an input to judge 'outlierness'.
- *Method output* is the output each method assigns to observations, binary ('outlier'/'inlier') or continuous ('outlierness').
- *Overall completeness* is the method's completeness of ability to identify outliers.

As non-labeled data is used in clustering (data without an attribute categorizing observations into 'outlier'/'inlier'), the thesis leaves out discussions about supervised scenarios for outlier detection.

2.2. LITERATURE REVIEW ON CLUSTERING

The literature review on clustering aims to discuss the objectives of clustering and how clustering results are evaluated. For objectives of clustering, literature is identified from several authors, providing their respective definition of objectives. The definition of objectives is discussed to create a unified definition, later quantified when discussing cluster evaluation.

For evaluation of clustering, quantification of clustering objectives is conducted with evaluation metrics. Advantages and disadvantages of evaluation methods are discussed along with a discussion of how different evaluation metrics have mutual dependence. Furthermore, a set of practical illustrations is provided.

2.3. APPLICATION OF OUTLIER DETECTION AND CLUSTERING

Finally, to quantify how outlier detection can improve clustering results, the following steps are executed:

- Step 1: Dataset is selected.

A wide range of datasets with different properties are selected for analysis. It is striven that datasets vary in number of attributes (variables) and observations, vary in composition of categorical and continuous attributes and vary in statistical characteristics. Furthermore, a discussion of selection of datasets is conducted in the analysis.

- Step 2: Outlier detection method is applied, and top-n % outliers are removed

For each dataset, prior to applying clustering algorithms, selected outlier detection methods are applied for removal of top-n outliers. For equal comparison across outlier detection algorithms, top 5%, 10% and 15% outliers are removed based on the algorithms 'outlierness' output for observations. Simultaneously, 5%, 10% and 15% random selected observations are removed, for later comparison of clustering results. Removal of random selected observations is conducted to evaluate whether outlier removal improves clustering results at all. That is, comparing apples with apples.

- Step 3: Clustering is applied to treated dataset and evaluated

For each treated dataset, clustering algorithms are applied, and results are discussed based on the clustering evaluation criteria identified. It is summarized how outlier detection and clustering algorithms excel in combination for datasets of different properties.

3. LITERATURE REVIEW

The term *outlier* was introduced in 1969 when Grubbs (Grubbs, 1969) argued that “*an outlying observation, or ‘outlier’, is one that appears to deviate markedly from other members of the sample in which it occurs*”. Following Grubbs (1969), Hawkins (1980) argued that “*an outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*” while Barnett and Lewis (1994) defined an outlier as “*an observation or subset of observations which appears to be inconsistent with the remainder*”. However, prior to Grubbs and Hawkins, Benjamin Pierce (1852) and Charles Pierce (1873) published papers respectively mentioning: “*Criterion for the Rejection of Doubtful Observations*” and “*On the Theory of Errors of Observations*”, discussing observations with deviating characteristics. Thus, the idea of observations with outlying characteristics has existed for a longer period.

There is not a single definition of an outlier, making it subject to discussion. However, there is a consensus that an outlier in some way has deviating characteristics. For the remaining thesis, an inlier is defined as the contradiction of an outlier, meaning that an inlier does not have deviating characteristics:

“*An outlying observation, or ‘outlier’, is one that appears to deviate markedly from other members of the sample in which it occurs*”
- Grubbs (1969)

“*An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*”
- Hawkins (1980)

“*An observation or subset of observations which appears to be inconsistent with the remainder*”
- Barnett & Lewis (1994)

“*An outlier is an observation distant to other observations*”
- Knorr & Ng (1997)

“*An outlier in a set of data is an observation or a point that is considerably dissimilar or inconsistent with the remainder of the data*”
- Ramaswamy, Rastogi, & Shim (2000)

“ An object is an outlier if it is in some way significantly different from its neighbors ”
- Papadimitriou, Gibbons, & Faloutsos (2003)

3.1. METHODOLOGY FOR IDENTIFICATION OF LITERATURE ON OUTLIERS

The discipline of outlier detection has been thoroughly researched. ‘Outlier detection’ generates 57.000 results on ScienceDirect, making it impossible to review the literature article by article. To create an overview of the scope of literature, following steps are executed:

1. The most well-cited meta studies (review articles) within the field of outlier detection are downloaded from ScienceDirect, based on the keyword ‘outlier’.
2. The outlier detection methods, from the review articles, are compared across articles to identify the most frequent reviewed methods, approaches and authors (domain experts).
3. The identified methods are compared with general search machine results and Wikipedia pages for outlier detection¹ and anomaly detection² to get a consensus of the identified literature.
4. The reviewed literature, identified in the review articles, is downloaded from ScienceDirect with the most recent first. For each article, a chapter describing the current literature until publication of the respective article, is inspected.
5. If the inspected chapter refers to literature not identified earlier, the newly identified literature is downloaded.

With the above-mentioned approach for literature identification, a ‘history’ of evolution in the discipline of outlier detection is mapped. The downside for this approach is the lack of recent literature as a result of missing review papers.

Following criteria and considerations are put in place while executing previously described process:

- The identified literature must be applicable to both discrete and continuous data.

This criterion is created to solely identify universal applicable methods, avoiding literature dealing only with certain types of data (e.g. time series or spatial data). The only exemption for this criterion is the

¹ <https://en.wikipedia.org/wiki/Outlier>

² https://en.wikipedia.org/wiki/Anomaly_detection

early literature of outlier detection (statistical test) which is included to create a fundamental understanding of outlier detection and its origin.

- The identified literature must be original research proposing an approach to outlier identification not previously seen

This criterion is created to avoid literature suggesting how to deal with outlier detection within a certain field (e.g. logistic systems, healthcare, fraud systems) by applying an already developed method to a practical case.

- The identified literature must be well cited (≥ 50) and/or preferably from a notable author within the discipline

This criterion is created to avoid spending a disproportionate amount of time compared to importance of the literature. Literature without important contributions to the discipline generally fails to be cited.

- The theory in the identified literature must have a continuous element

This criterion is created with respect to the methodology of removing top-n % of outliers based on outlier method output. The identified literature can have a binary output, as a threshold for observations to be identified as outliers, if the threshold is based on an underlying continuous element returning outlierness for observations.

3.2. OUTLIER DETECTION APPROACHES

Table 1 highlights 29 identified crucial papers. The literature is divided into approaches and shows the number of citations per paper. The number of citations is a conservative estimate based on numbers from the source, mostly being ScienceDirect or ResearchGate.

As with many fields of science, the approaches are characterized by one or more authors introducing a new approach to a problem, followed by other researchers adopting the approach for improvement of the theory. Literature dealing with distance-based or more complex approaches tends to propose a new algorithm for outlier detection, followed by one or more papers suggesting a modified algorithm for computational optimization.

With respect to the research question and the limitations stated earlier in this thesis, the literature solely dealing with computational optimizations is not discussed in this literature review.

Authors	Paper name	Approach	Citations
Grubbs (1950)	Sample Criteria for Testing Outlying Observations	Statistical test	644
Wilk, Gnanadesikan (1968)	Probability Plotting Methods for the analysis of data	Statistical test	728
Grubbs (1969)	Procedures for Detection Outlying Observations in Samples	Statistical test	1575
Gnanadesikan, Kettenring (1972)	Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data	Statistical test	558
Tukey (1977)	Explorative Data Analysis	Statistical test	16676
Garrett (1989)	The chi-square plot: A tool for multivariate outlier recognition	Statistical test	61
Iglewicz, Hoaglin (1993)	How To Detect and Handle Outliers	Statistical test	580
Filzmoser, Garrett, Reimann (2005)	Multivariate Outlier Detection in Exploration Geochemistry	Statistical test	190
Vanderviere, Huber (2008)	Adjusted Boxplot for Skewed Distributions	Statistical test	169
Knorr, Ng (1997)	A Unified Approach for Mining Outliers	Distance-based	57
Knorr, Ng (1998)	Algorithms for Mining Distance-Based Outliers in Large Datasets	Distance-based	979
Knorr, Ng (1999)	Finding Intensional Knowledge of Distance-Based Outliers	Distance-based	310
Ramaswamy, Rastogi, Shim (2000)	Efficient Algorithms for Mining Outliers from Large Data Sets	Distance-based	921
Angiulli, Pizzuti (2002)	Fast Outlier Detection in High Dimensional Spaces	Distance-based	235
Hautamaki, Kärkkäinen, Fränti (2004)	Outlier Detection Using K-nearest Neighbour Graph	Distance-based	50
Fan, Zaiane, Foss, Wu (2006)	A Nonparametric Outlier Detection for Effectively Discovering Top-N Outliers from Engineering Data	Distance-based	56
Zhang, Hutter, Jin (2009)	A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data	Distance-based	198
Breuning, Kriegel, Ng, Sander (2000)	LOF: Identifying Density-Based Local Outliers	Density-based	1741
Jin, Tung, Han (2001)	Mining Top-n Local Outliers in Large Databases	Density-based	215

Authors	Paper name	Approach	Citations
Tang, Chen, Fu, Cheung (2002)	Enhancing Effectiveness of Outlier Detections for Low Density Patterns	Density-based	146
Papadimitriou, Kitagawa, Gibbons, Faloutsos (2003)	LOCI: Fast Outlier Detection Using the Local Correlation Integral	Density-based	482
Jin, Tung, Han, Wang (2006)	Ranking Outliers Using Symmetric Neighborhood Relationship	Density-based	139
Latecki, Lazarevic, Pokrajac (2007)	Outlier Detection with Kernel Density Functions	Density-based	84
Kriegel, Kröger, Schubert, Zimek (2009)	LoOP: Local Outlier Probabilities	Density-based	115
Schubert, Zimek, Kriegel (2014)	Generalized Outlier Detection with Flexible Kernel Density Estimates ³	Density-based	44
Lazarevic, Kumar (2005)	Feature Bagging for Outlier Detection	High-dimensional	187
Kriegel, Schubert, Zimek (2008)	Angle-Based Outlier Detection in High-Dimensional Data	High-dimensional	150
Kriegel, Kröger, Schubert, Zimek (2009)	Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data	High-dimensional	68
Kriegel, Kröger, Schubert, Zimek (2011)	Interpreting and Unifying Outlier Scores	High-dimensional	60

Table 1: Identified literature

For the identified literature a trend is observed. Generally, the early literature mentions outliers in a statistical testing context, where the discipline of outlier detection serves the purpose of removing outliers to improve hypotheses testing and statistical significance. Early literature perceives outlier detection as the probability of an observation belonging to a distribution, whereas more recent literature mentions outliers in a context of discovering anomalies in large datasets. Furthermore, more recent literature is largely divided in to distance-based, density-based and high-dimensional approaches. Common for this literature is the abandonment of probability theory in favour of multidimensional

³ Included although citations <50 as Kriegel and Schubert are considered domain experts with previous papers

proximity. Thus, with the improvement and adoption of computational power, scope has moved from statistical testing, mostly a univariate discipline, to data mining, a multivariate discipline.

To have a unified approach and definition of outlier detection for the remaining literature review, outlier detection is defined as follows:

- *Outlier detection is the discipline of identifying observations with deviating characteristics, by the means of different approaches (e.g. statistical test, density-based approach).*

3.2.1. Statistical Z-test

Statistical test for outliers assumes an underlying normal distributed population (Grubbs, 1969), also called parametric test, meaning that a set of parameters are assumed fixed. Non-parametric test differs from parametric test by having a set of parameters that can vary in numbers and does not necessarily rely on an assumed underlying distribution.

Grubbs argued that statistical outliers should be identified as observations that “*would exceed the ‘critical value’ based on random sampling theory*”, and that critical value “*is generally recommended [...] a low significance level, such as 1%*”.

A simple univariate approach to test if observations exceed a critical value is the Z-score:

$$Z_i = \frac{X_i - \bar{X}}{SD}$$

where $\bar{X}$ is the sample mean and SD is the sample standard deviation.

The Z-score method is based on the idea that X follows a normal distribution, $N(\mu, \sigma^2)$, and Z-scores follow a standard normal distribution, $N(0,1)$. The Z-score reflects the number of sample standard deviations observation X_i differs from $\bar{X}$. An outlier is identified as $|Z| > 3$, based on the idea that 99.7% of the observations are within 3 standard deviations from the mean in a normal distribution and anything greater would be rejected as an outlier belonging to a different distribution.

However, the Z-score, using the standard deviation, lacks robustness. As the standard deviation can be affected by extreme values, meaning that observations that are considered outliers in a sample, are identified as inliers, when an even more extreme observation enters the sample. Thus, the standard deviation possesses a weakness, as it is a measure of deviation in a sample, based on the sum of deviations from the mean including the very outlying observations the user is trying to identify.

3.2.2. Robust statistics

To overcome this issue, Iglewicz & Hoaglin (1993) suggested an approach based on *robustness*. Robust statistics focus on the median as a measure of central tendency, while the mean is viewed as a measure vulnerable to extreme values. To replace the standard deviation, Iglewicz & Hoaglin (1993) introduced the absolute deviation of the median (MAD):

$$MAD = median(|X_i - \bar{X}|)$$

where $\bar{X}$ is the sample median.

The modified Z-score approach suggested is:

$$M_i = \frac{0.6745(X_i - \bar{X})}{MAD}$$

where 0.6745 is equal to the number of standard deviations having 50% of the observations within in a normal distribution. It was suggested that observations with $|M_i| > 3.5$ should be identified as outliers.

3.2.3. Tukey's boxplot

With his book *Explorative Data Analysis*, Tukey (1977) presented a visual approach for exploring data and provided the boxplot as a simple approach to display properties of a distribution. Having no underlying assumptions of a normal distribution, the method is non-parametric. Tukey's boxplot has since then been widely accepted as a default method to identify outliers in univariate data.

Tukey argued that outliers are identified using inner and outer fences:

$$\text{lower inner fence} = Q_1 - 1.5(Q_3 - Q_1) \quad \text{and} \quad \text{upper inner fence} = Q_3 + 1.5(Q_3 - Q_1)$$

$$\text{lower outer fence} = Q_1 - 3(Q_3 - Q_1) \quad \text{and} \quad \text{upper outer fence} = Q_3 + 3(Q_3 - Q_1)$$

where Q_1 and Q_3 are the first and third quantiles respectively, and $Q_3 - Q_1$ is the interquartile range (IQR). Any observation between the inner and outer fences are identified as moderate outliers, and observations lower than the *lower outer fence* or greater than the *upper outer fence* are identified as severe outliers.

Tukey's boxplot (Tukey, 1977) has great properties in univariate data, being a robust statistical tool based on quartiles while graphically reflecting the characteristics of a distribution such as distance between mean and median, length of tails, skewness and outliers. Although Vanderviere & Huber (2008) criticized

Tukey's boxplot (Tukey, 1977) for lack of properties to deal with skewed data, leading to the adjusted boxplot, the boxplot approach has its greatest shortcomings in multivariate data.

3.2.4. Mahalanobis distance

The scatterplot exhibits the observations of two variables (X, Y) in a two-dimensional space. When a point deviates significantly from the cloud of points of the other observations, it may indicate an outlier. To quantify its deviation, Euclidean distance is useful to derive the distance from the center of gravity (0, 0), to the atypical observation. Euclidean distance is defined as:

$$D(q, p) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$

where q_i and p_i are the coordinates of observations and i is the number of the respective dimension. Thus, to calculate the distance to the center of gravity for a standardized dataset, where center of gravity is given by $\vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, the coordinates are squared, summed and square rooted.

However, applicability of Euclidean distance in correlated data has weaknesses. For example, in Figure 1, it identifies observation A and B with a distance from the center of gravity of 2.76 and 3.88, respectively, expressing greater outlierness for observation B than observation A. That is despite of observation B following the correlated pattern of the data, having less deviating characteristics than observation A.

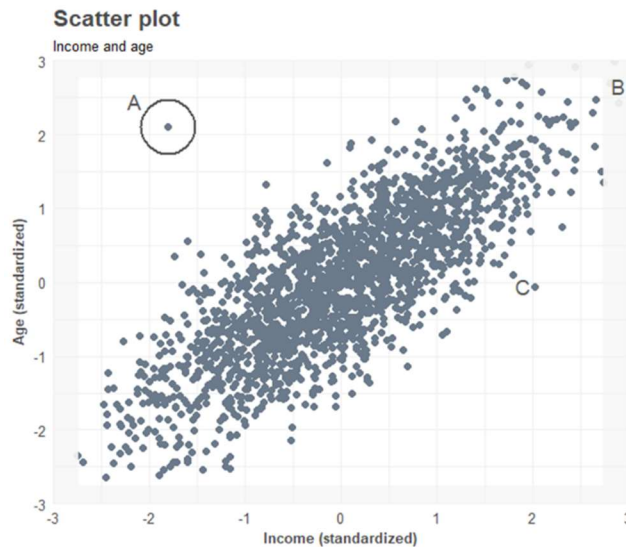


Figure 1: Scatterplot for variables income and age

To deal with outliers for regression and multivariate data, it is necessary to take the covariance of the variables into account. For that, Mahalanobis distance (Mahalanobis, 1936) is useful. Mahalanobis distance identifies a set of vectors, corresponding to the number of dimensions, with the first vector explaining the most variance, which in this case is the vector fitting the correlated pattern. The second vector explains the second most variance and is perpendicular with the first vector.

Mahalanobis distance is defined as:

$$MD(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

where $\vec{x}$ is a vector of observations with mean values given by the vector $\vec{\mu}$ and S^{-1} is the inverse covariance matrix for variables.

For the data in Figure 1, a covariance matrix $S = \begin{pmatrix} 1.00 & 0.80 \\ 0.80 & 1.00 \end{pmatrix}$ express a correlation of .80 between the two variables. For demonstration purpose, Mahalanobis distance is visualized in Figure 2.

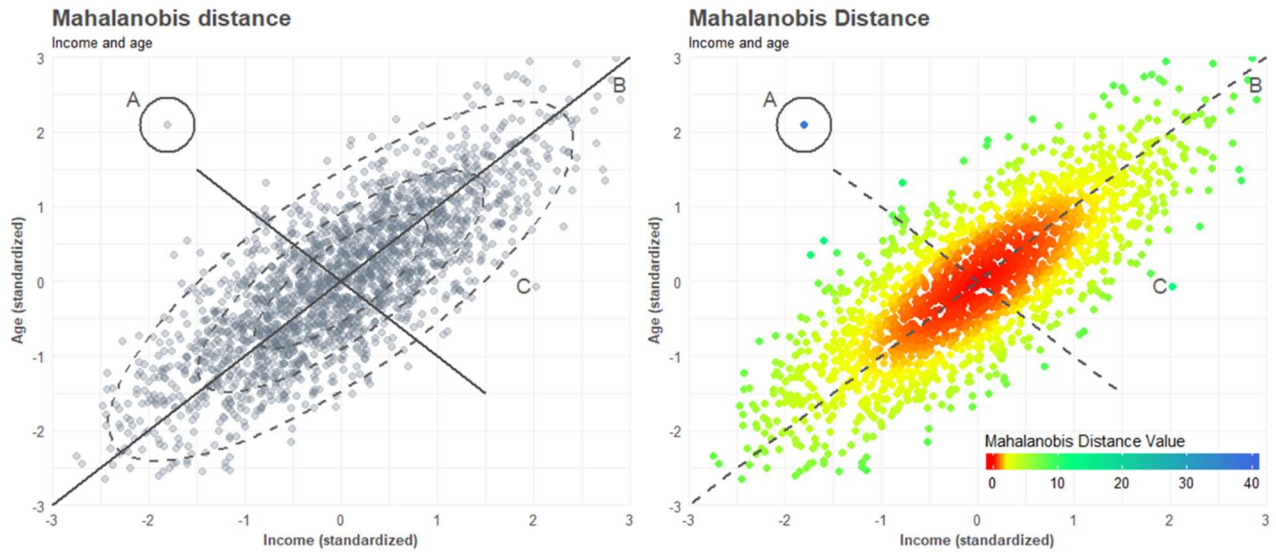


Figure 2: Mahalanobis distance visualized with standard deviations (left) and value (right)

Inspecting Figure 2, the standard deviation for the vectors follows the correlated pattern of the data, meaning that 68-95-99.7 confidence intervals, with 1, 2 and 3 standard deviations respectively, is given by an ellipse shape. Observations deviating on the second vector have a greater distance from the center of gravity, as they deviate the correlated pattern of the data.

Observation B is considered the most outlying observation from center of gravity based on Euclidean distance, but does not differ substantially from observation C with Mahalanobis distance. Observation A is by far to be considered the observation with the most outlying characteristics following Mahalanobis distance. Although using Mahalanobis distance for outlier detection is not based on a certain threshold, a probability density plot by Mahalanobis distance will show the observations deviating greatest in the tail.

Mahalanobis distance is applicable to datasets characterized by continuous data with collinearity and should mostly be used within this specific domain. As with univariate data, usage of robust estimates for mean and standard deviation in Mahalanobis distance, will perform better when working with correlated data with extreme outliers. Working with uncorrelated data, non-robust Mahalanobis distance will perform worse than Tukey's boxplot, considering that Tukey's boxplot is robust, hence not being sensitive to extreme outliers. Furthermore, Mahalanobis distance will fail when working with categorical variables, as Mahalanobis distance is based on covariance and means.

3.2.5. Q-Q plot and Chi-square plot

The Q-Q plot (quantile-quantile plot), was suggested (Wilk & Gnanadesikan, 1968) as a tool for investigating if a sample distribution follows a normal distribution. The Q-Q plot has versions with different properties, the most popular being the version plotting a given sample distribution against a theoretical standard normal distribution $N(0,1)$. For that version, data is ordered in increasing order ($X_1 < X_2 < \dots < X_i$), and a standard normal distribution is divided in to $n + 1$ areas, with n being the number of observations in a sample, with each area having equal probability density. The i th ordered observations are in increasing order plotted against the $\frac{i}{n+1}$ th quantile for a standard normal distribution, reflecting if the sample follows a normal distribution.

The Q-Q plot can technically plot any sample distribution against another distribution, making it a great tool to discover outliers in the tails of a sample distribution, if plotted against an appropriate distribution. The Q-Q plot has its strength in univariate data and for multivariate data with a maximum of two dimensions. As it requires a pairwise inspection, variable by variable, it is time consuming for multidimensional data (>2d) and to some extent also subjective.

For a more effective outlier detection method following same philosophy of quantile inspection, Garrett (1989) suggested to use the chi-square plot. The chi-square plot plots the squared Mahalanobis distance against the quantiles of a chi-square distribution (X_p^2), with p degrees of freedom, allowing the user to

inspect a multivariate distribution against a theoretical distribution, leading to identification of observations deviating from multivariate normality in the tails. The procedure allows the user to delete the most extreme observations, until the observations fall into a straight line. This method is more effective than pairwise inspection of multiple distributions, as it essentially is a multidimensional Q-Q plot.

3.2.6. Distance-based approaches

Going from statistical testing to distance-based approaches for identifying outliers, scope in the literature changes from probabilities of an observation belonging to a distribution, to knowledge discovery in large datasets, also called data mining. With different philosophy of distance and deviation, data mining methods leave out centrality and dispersion measures such as mean, quantiles and standard deviation, characterized by statistical testing.

Ramaswamy, Rastogi, & Shim (2000) argued that the main problem with statistical testing is that *'the user might not have enough knowledge about the underlying data distribution'* and the main benefit of distance-based approaches is that *'it does not require any apriori knowledge of data distributions that the statistical methods do'*.

Knorr & Ng (1997) argued that despite of all the statistical test available for outlier detection, *'there is no guarantee of finding outliers'* as *'there might not be any test developed for a specific combination or because no standard distribution can adequately model the observed distribution'*. Thus, the desire with distance-based approaches is to replace assumptions of underlying data distributions with observed data distributions.

3.2.6.1. DB Outlier

Knorr & Ng (1997) introduced distance-based approaches by focusing on neighborhood of observations. The distance-based methodology was defined as *'An object O in a dataset T is a $UO(p, D)$ -outlier if at least fraction P of the objects in T are $\geq$ distance D from O '*. They suggested a rather simple method called DB-outlier.

Let the distance D , which can take any distance input, be a radius from object O in a two-dimensional space. The distance D is set to 1.5. Let the fraction parameter P , which can take any input between 0 and 1, be set to 0.05 or 5% of dataset T . Let observations within object O 's neighborhood, given distance measure D , be labeled as observations Q_n . If the proportion of Q_n within observation O 's neighborhood,

over the total number of observations in dataset T , is less than the fraction parameter P , observation O is considered an outlier and vice versa. Outliers are then given by:

$$Outliers(D, P) = \frac{(Q | dist(O, Q) < D)}{T} \leq P$$

Inspecting Figure 3, which was created for demonstration purposes, the distribution for dimension 1 and 2 is not normally distributed, but rather bimodal. The observation O , of which DB-outlier method is demonstrated, would hide in the bimodal peaks in the distributions for dimension 1 and 2. For left plot, observation O has a neighborhood given by distance D , in which two neighboring observations Q_n are identified. The proportion of neighboring observations Q_n is less than fraction parameter P of 0.05, or 5% of data. Observation O is therefore identified as an outlier.

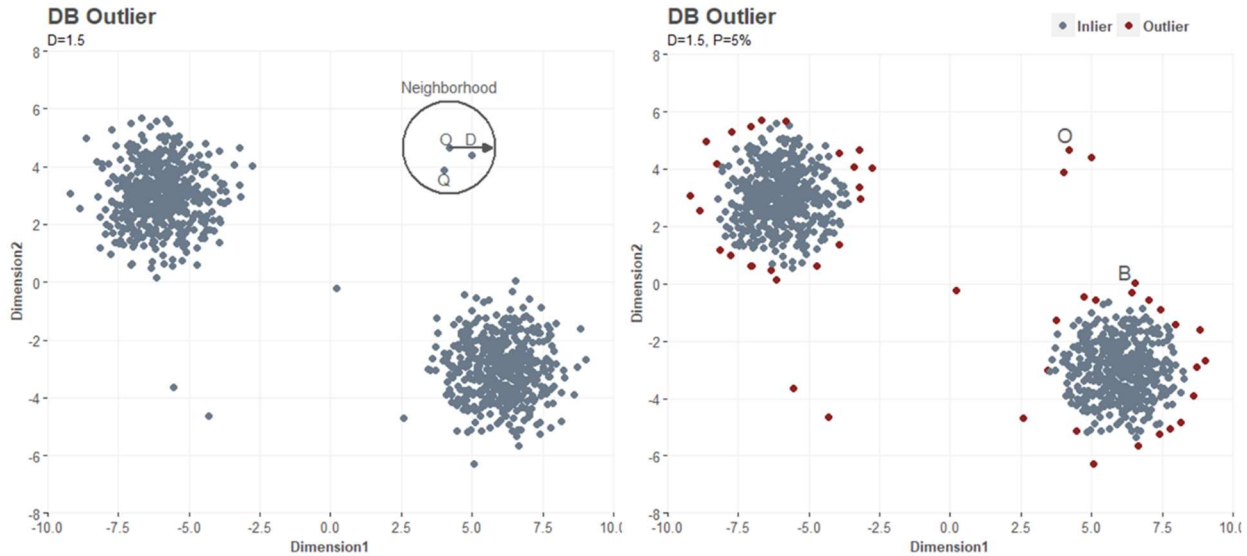


Figure 3: Distance-based outlier approach (left) and result (right)

Where statistical test is based on assumptions of certain underlying distribution and centrality, leading to lack of properties to deal with high-dimensional data, the distance-based approach DB-outlier, performs well to identify anomalies in datasets characterized by many dimensions and non-normal distributions. Being the first approach suggested for outlier detection within the discipline of data mining, the model builds a solid foundation based on concept of distance and neighborhood.

The theory is however not completely fit for outlier detection in multivariate data, as it has two major shortcomings. The first shortcoming is the binary output. Observation O and observation B (see: Figure 3) are both identified as outliers, but do not have the same deviating characteristics. Observation B is

close to one of the two gaussian clusters, while observation O is distant to other observations. Thus, the binary output does not provide the user with an idea of outlierness for observations. The second shortcoming is the subjective threshold of parameter P and the distance D , leaving the user with no options but to loop over different input parameters. A large distance D and low P will only identify the most deviating observations, whereas small distance D and high P will identify more observations as outliers. A solution to both shortcomings is to omit parameter P , and let the user judge each observation by the proportion of observations within its neighborhood, over the total number of observations, providing an output between 0 and 1.

3.2.6.2. K Nearest Neighbor

As with a lot of pioneering research, the DB-outlier method is the first within its field, but others adopted the idea of distance and neighborhood to improve the theory.

Ramaswamy, Rastogi, & Shim (2000) suggested an algorithm called 'k-NN', based on the distance to k nearest neighbor for each observation p in dataset. The distance measure for observation p to k nearest neighbors is noted as $dist^{kNN}(p)$. The idea is that observations with low $dist^{kNN}(p)$ have a dense neighborhood, meaning that observation p does not have deviating characteristics. Top n observations with high $dist^{kNN}(p)$ have sparse neighborhoods and deviating characteristics and will be considered outliers.

The k-NN approach is intuitive and provides a measure of 'outlierness' allowing the user to judge how deviating an observation is, hence avoiding binary output. Furthermore, different measures of distance can be taken, such as Euclidean distance and Manhattan Distance, increasing the flexibility of the algorithm.

However, as well as it can be an advantage to have continuous output and many input options, it also leaves the user with a subjective decision making. The subjective difficulties consist of setting a threshold of $dist^{kNN}(p)$ for top n observations to be identified as outliers and setting a factor of k .

For demonstration purposes, Figure 4 is created. The left plot has $k = 1$ and right plot has $k = 10$. With $k = 1$, each observation p will only return distance to nearest observation, failing to identify a group of close observations in an outlying cluster. The three observations seen in between the two major gaussian clusters of observations in Figure 4 are therefore identified as inliers. If k is large (such as $k = 50\%$ of n), it will produce less variance in the output vector of $dist^{kNN}(p)$. Less variance equals more similar $dist^{kNN}(p)$ for observations, making it more difficult to identify the outliers with the strongest

outlierness. This leads to a trial and error process of iterations over different factors for k , further increasing subjectivity when eventually choosing threshold.

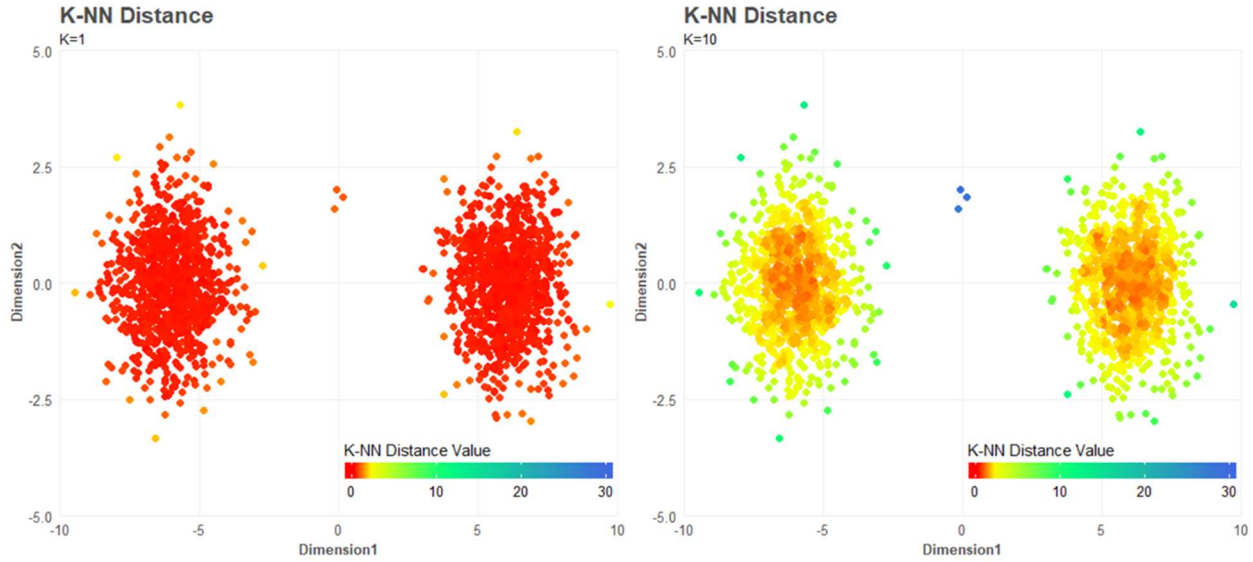


Figure 4: K-NN distance for $k=1$ (left) and $k=10$ (right)

To limit the issue of subjectivity, Angiulli & Pizzuti (2002) adopted the k-NN methodology but suggested an algorithm aggregating the k-NN with k as an input range. With $k = \{4, 5, \dots, 10\}$, the results of 4NN, 5NN, ... and 10NN is aggregated. The aggregated measure is a weight and is given by:

$$w_k(p) = \sum_{i=1}^k dist(p, nn_i(p))$$

where nn_i is the i th nearest observation/neighbor of observation p , given parameter k .

The algorithm leaves the user with less subjectivity and it gives a simpler output and smoother result of outlierness for observations.

Hautamaki & Ismo (2004) adopted the k-NN theory and proposed an algorithm combining network theory and k-NN, naming it k-NN Graph. It was suggested to calculate the in-degree for each observation, given k nearest neighbors. An observation with sparse neighborhood has low in-degree as few other observations connects to it due to missing proximity, and vice versa. If the in-degree for an observation is less than a threshold T , it is considered an outlier.

Inspecting Figure 5, which was created for demonstration purposes, left plot shows the links of observations identifying neighbors by 'linking'. With $k=3$, observation A connects to its three nearest

neighbors, being observation B, C and D, giving each observation one in-degree. The closest observation in the neighborhood of observation A, is observation B. Observation B connects to observation C, D and E as the closest neighboring observations, leaving no in-degree for observation A. The same accounts for observation C and D, meaning observation A is not in the neighborhood of any observation when $k=3$. Thus, observation A is identified as an outlier, when the threshold T is set to ≤ 2 .

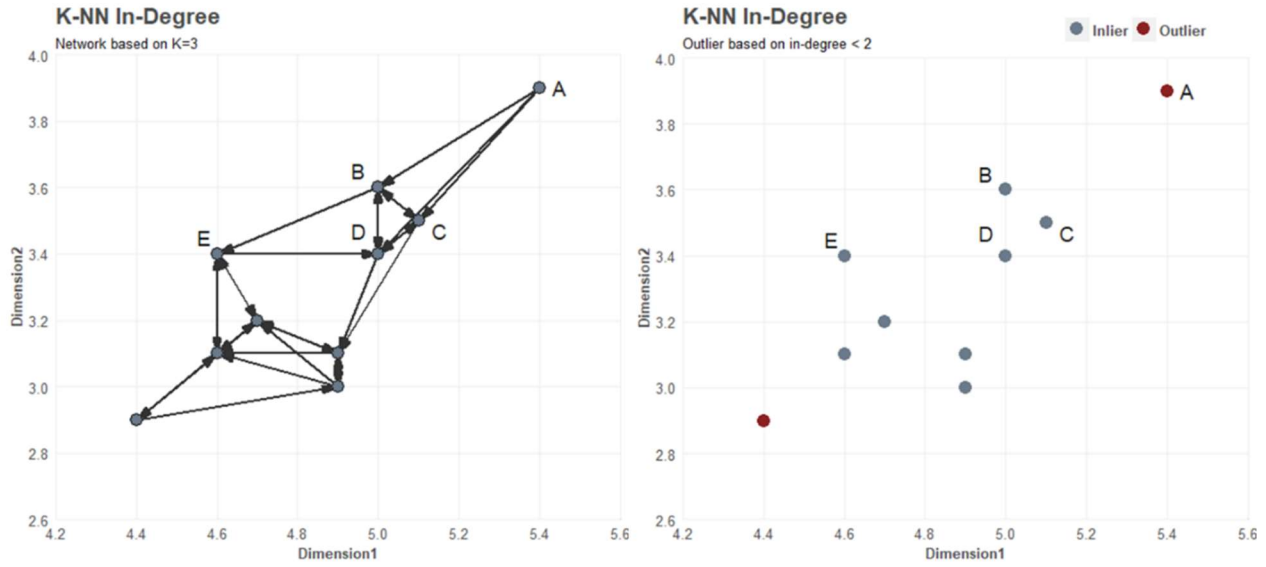


Figure 5: K-NN Graph for synthetic data with network (left) and classification (right)

As with DB-outlier, the output for K-NN Graph is a binary label. Again, it is concluded that a small k will fail to identify an outlying cluster of observations, as the observations in the given cluster simply would connect to each other with in-degrees. Although not suggested by Hautamaki et al. (2004), the eigenvector centrality metric from network theory is applicable to overcome this issue. The eigenvector centrality returns a number indicating how well-connected an observation is to other well-connected observations, exposing an outlying cluster of observations with low in degree values. Again, it is concluded that the threshold should be removed if the user wants to inspect the outlierness for each observation to provide an idea of how deviating observations are.

3.2.7. Density-based approaches

Density-based approaches were introduced by Breunig, Kriegel, Ng, & Sander (2000), following the introduction of distance-based approaches. As with distance-based approaches, it was argued that outlier detection based on statistical test has major drawbacks in data mining as most methods are univariate and the underlying distribution often is unknown. However, it was also argued that the distance-based

approaches lacked properties to deal with neighborhoods of different densities, due to being a ‘global’ outlier detection method. A ‘global’ methods means the entire dataset is taken in to account when comparing observations for outlierness. For DB-outlier, the amount of observations in the neighborhood given by distance D , is compared to the total amount of observations in dataset T . For various variants of k-NN, distance is compared to the distance to nearest neighborhood for every other observation in dataset. Thus, the global methods lack properties to identify outliers in neighborhoods of different densities.

Inspecting Figure 6, which was created for demonstration purposes, it is observed that observations in gaussian cluster $C1$ have a dense neighborhood and are identified as inliers while observations in cluster $C2$ are identified as outliers although the observations do not have deviating characteristics, *relatively to their neighbors*, and although $P = 0.01$, meaning that just 1% of observations within a distance of 3 is necessary to be an inlier.

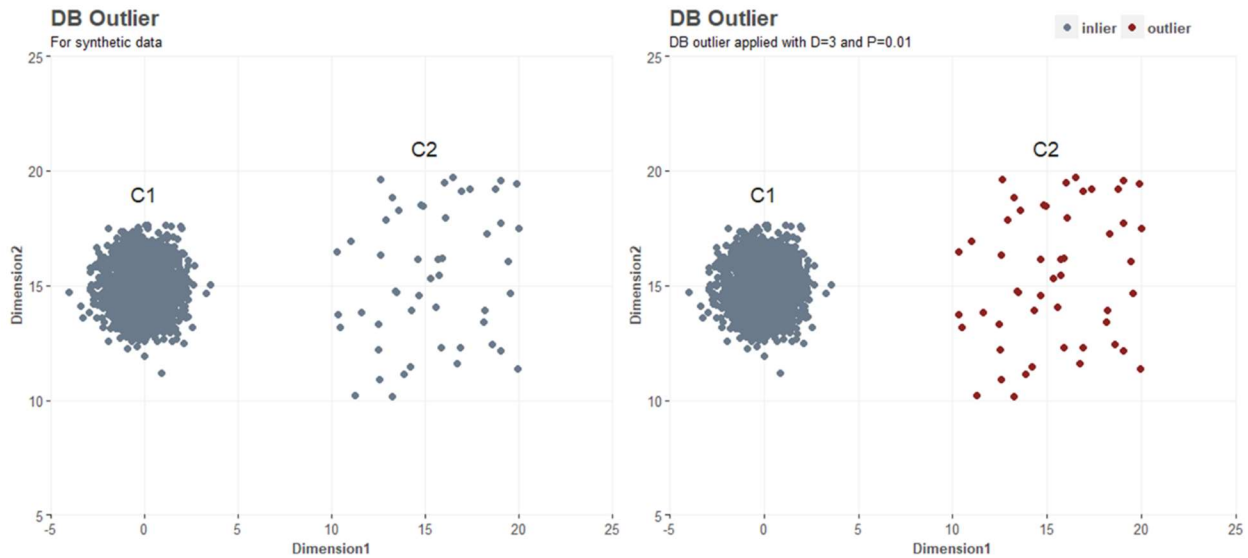


Figure 6: Synthetic data (left) taken from Breuning et. al (2000) with DB outlier applied (right)

For an analogy with income and age, the philosophy of outlier detection for density-based approaches is to compare Warren Buffet with Bill Gates, Carlos Slim and other billionaires, leaving out global comparison with 99% of the remaining population. As an example, Mark Zuckerberg aged 32 is an outlier *relatively* to other billionaires, such as Warren Buffet, Bill Gates and Carlos Slim with an age of 86, 61 and 77, respectively.

3.2.7.1. Local Outlier Factor (LOF)

With more than 1.700 citations, Breunig, Kriegel, Ng, & Sander (2000) successfully introduced density-based approaches as a new domain in outlier detection. To identify outliers locally, it was argued that observations outlierness should be perceived relatively to the neighborhood, leaving out the global data distribution. The intuition is that outliers are identified as having different density in the neighborhood, when compared to the density in the neighborhood of neighboring observations.

An algorithm was proposed, taking only input parameter k , being the number of nearest neighbors to compare neighborhood density with. For each observation o in dataset, the neighborhood boundaries, given parameter k , are defined as:

$$reachDist_k(o, p) = \max\{kDist(p), d(o, p)\}$$

meaning the most distant k_{th} observation defines the boundaries of the neighborhood. As two or more observations can have equal distance to observation o , the number of observations within the neighborhood can be greater than k .

The density of the neighborhood is given by the local reachability density computed as:

$$lrd_k(o) = 1 / \left\{ \frac{\sum_{p \in N_k(o)} reachDist_k(o, p)}{|N_k(o)|} \right\}$$

where $|N_k(p)|$ is the number of observations within the neighborhood, given a certain k . Thus, the local reachability is 1 over the accumulated distance from neighboring observations to observation o , over the number of observations in the neighborhood of observation o .

The local outlier factor (LOF) is then computed as:

$$LOF_k(o) = \frac{\sum_{p \in N_k(o)} \frac{lrd_k(p)}{lrd_k(o)}}{|N_k(o)|} = \frac{\sum_{p \in N_k(o)} lrd(p)}{|N_k(o)|} / lrd_k(o)$$

which is the average local reachability density for observations $\{p_1, p_2, \dots, p_k\}$ in the neighborhood of observation o , divided by the local reachability density for observation o . A LOF close to 1 indicates equal neighborhood density relatively to neighboring observations, whereas a LOF significantly lower (e.g. 0.5) indicates relatively denser neighborhood. Significantly higher LOF-score, say 5, indicates relatively sparse neighborhood and outlierness.

For demonstration purposes, the left plot in Figure 7 shows a sparse neighborhood for observation o , when $k = 3$. Observation p_1, p_2 and p_3 are closest neighbors to o , and all have denser neighborhoods, reflected by the size of the dashed *reachDist* for each p . The right plot shows two odd distributions, with small clusters of different densities. LOF is applied with $k = 10$ and a LOF score is assigned to observations. Observations are identified with sparse neighborhood, relatively to neighbors, with LOF-scores >3 . Thus, with its relative density methodology, LOF performs well in neighborhoods of different densities and provides a continuous scoring output, enabling users to identify the observations with strongest outlierness.

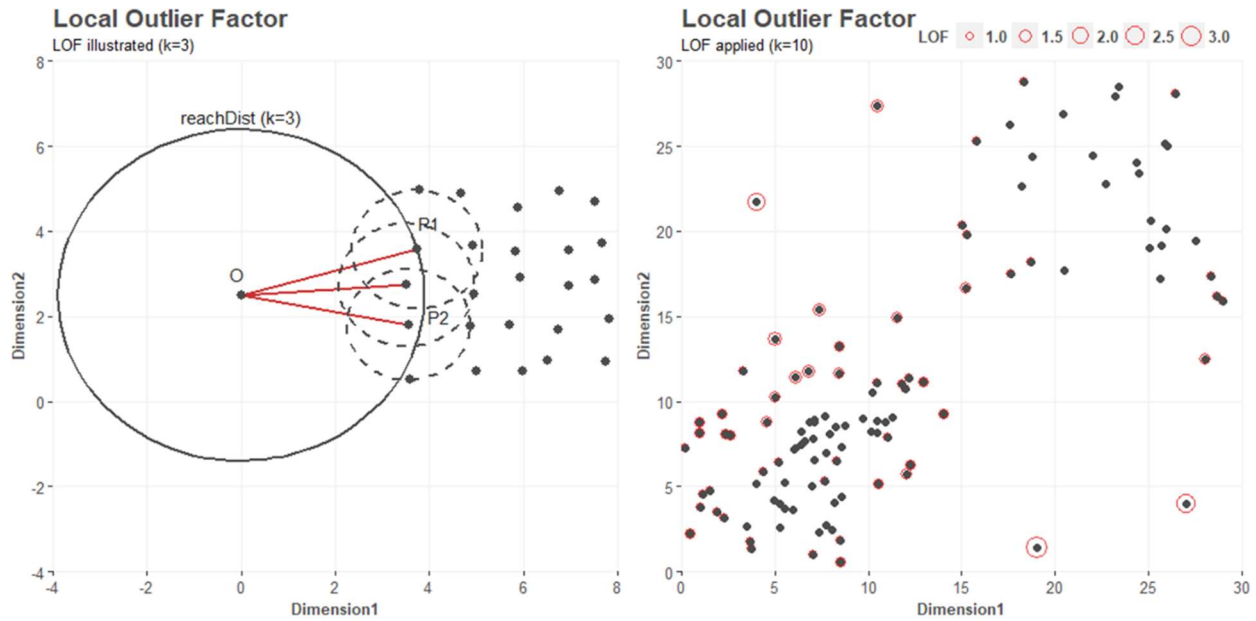


Figure 7: LOF illustrated (left) with example (right) showing LOF-score for observations

Density-based approaches was adopted (Jin, Tung, & Han, 2001; Jin, Tung, Han, & Wang, 2006), with the former concerning computational performance and the latter dealing with close clusters of different densities. The latter is of special interest as it is a contribution to the theory, optimizing outlier identifying for observations close to clusters of different densities. Jin, Tung, Han, & Wang (2006) criticised standard LOF by arguing that observations belonging to a sparse cluster, while being close to a dense cluster, will identify neighbors in the dense cluster for comparison. With neighbors in the dense cluster, the observation belonging to a sparse cluster will wrongly have a high outlier score assigned.

For demonstration purposes, the left plot in Figure 8 contains observation q as a local outlier and observaiton p as an inlier in a sparse cluster. The plot reveals observation p as identifying three closest

neighbors in the dense cluster. The right plot shows a LOF-score of 1.46 for observation p as it compares neighborhood density with observations in the dense cluster. Thus, observation p is identified as the second most outlying observation, with only observation q having greater outlierness. Observation p does not have deviating characteristics and is therefore incorrectly assigned with relatively great outlierness. To overcome this shortcoming, increasing k can be a solution as well as it can further increase the problem. When increasing k , users do not know if bordering observations identifies more neighbors in different cluster with denser neighborhood or identifies neighbors in the cluster it belongs to.

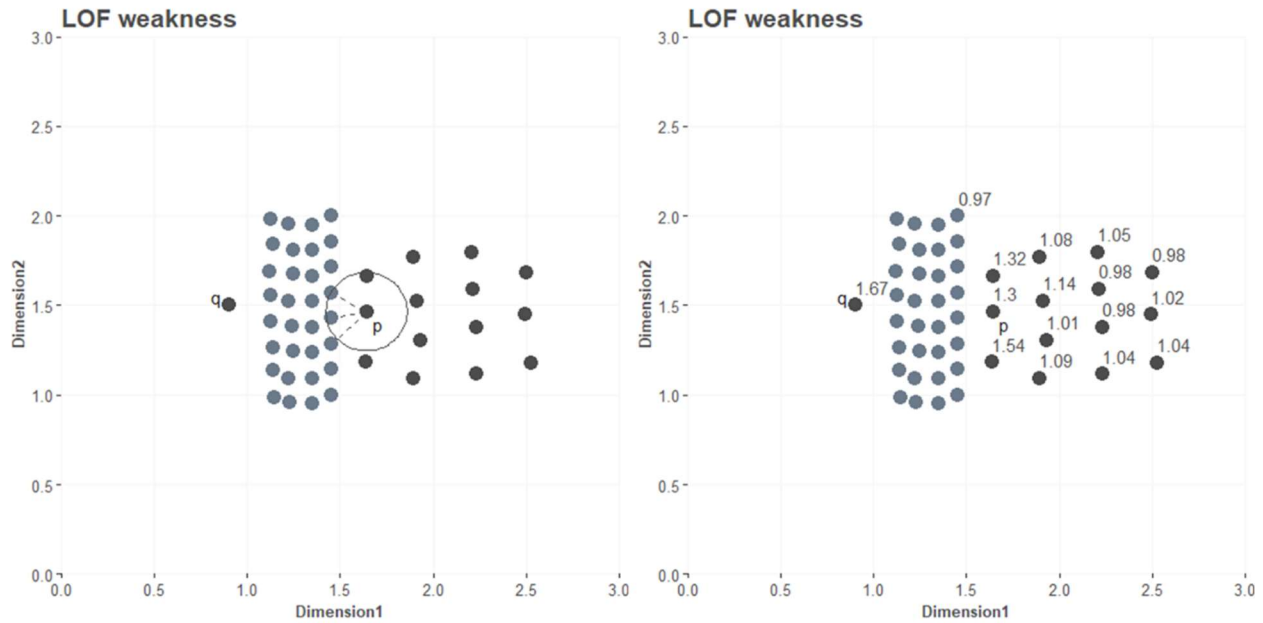


Figure 8: Two close clusters in black and blue of different density (left) and LOF-score for observations (right)

3.2.7.2. Influenced Outlierness (INFLO)

For the above-mentioned shortcoming, an algorithm called *influenced outlierness* (INFLO) was proposed (Jin et al., 2006). The algorithm takes symmetric neighborhood relationship in to account and considers an *influenced space* by the means of k-NN and the reverse nearest neighbor (RNN). The reverse nearest neighbor is described as an *inverse relation* in which other observations have spatial proximity to the observation subject to outlier identification. For simplification, we can return to Figure 5 demonstrating k-NN in-degree as suggested by Hautamaki & Ismo (2004). Observation A identifies observation B, C and D as neighbors when $k = 3$, but observation B, C and D does not identify observation A as a nearest neighbor. Thus, observation A is a reverse neighbor of observation B, C and D.

Applying the analogy to the demonstrated case in Figure 8, observation p has an inverse relationship with other observations in the sparse cluster, as these observations identify observation p as a nearest neighbor. For a set of observations $\{q_1, q_2, q_3, \dots, q_n\}$ with an inverse relationship to observation p , the nearest neighbor of those will be identified, avoiding impact from the actual nearest neighbors. The *influenced space* for observation p is then defined as $IS_k(p) = \{q_n, q_n, q_n\}$ where k is the nearest neighbors parameter for the set of inverse relation to observation p . The average density of the influenced space for observation p is given by:

$$den_{avg}(IS_k(P)) = \frac{\sum_{o \in IS_k(p)} den(o)}{|IS_k(P)|}$$

where $den(o)$ is the density of the neighborhood for observations with inverse relationship $\{q_1, q_2, q_3, \dots, q_n\}$ to observation p and $|IS_k(P)|$ is the number of nearest neighbors. INFLO-score is then given by:

$$INFLO_k(p) = \frac{den_{avg}(IS_k(P))}{den(p)}$$

Observations with INFLO-score around 1 have similar neighborhood density, when compared with neighborhood density in the influenced space. Observations with INFLO-score greater than 1 have less dense neighborhood relatively to neighborhood density in the influenced space, indicating outlierness for observation p . Thus, INFLO and LOF have similar output properties for scoring outlierness.

3.2.7.3. Connectivity-based Outlier Factor (COF)

J. Tang, Chen, Fu, & Cheung (2002) adopted the density-based approach and argued that measuring change in density may rule out outliers close to some non-outliers patterns that have low density. That is, outliers are difficult to detect in areas of low density, and choosing a large k may result in entire clusters of low density to be identified as outliers. To overcome this issue, J. Tang, Chen, Fu, & Cheung (2002) proposed to measure *isolativity* instead of *low density*, when measuring outlierness. For measuring isolativity, density is simply redefined as the chaining-distance between observations in the neighborhood. For a chaining-distance, each observation p_i has a set of nearest neighbors, denoted as $N_k(p_i)$. The number of nearest neighbors is given by k and is an input parameter. The nearest neighbors to p_i are ordered in ascending order, based on distance, denoted as $SBNpath$.

For each nearest neighbor in $SBNpath_i$, the trail to the next nearest neighbor is denoted as $SBNtrail_i$ and the distance for the $SBNtrail_i$ is a cost function, denoted as c_i . The average chaining distance for observation p_i is given by:

$$acDist_i = \sum_{i=1}^{r-1} \frac{2(k+1-i)}{k(k+1)} * dist(e_i)$$

where k is the number of nearest neighbors, i is the number of the i th observation and $dist(e_i)$ is the distance between $i-1$ th and i th observation for the $SBNtrail_i$.

The connectivity-based outlier factor for p_i is then given by:

$$COF_k(p_i) = \frac{|N_k(p)| * acDist_i}{\sum_{o \in N_k(p)} acDist_o}$$

where $\sum_{o \in N_k(p)} acDist_o$ is the sum of average chaining distance for observations in the nearest neighbor set of p_i and $|N_k(p)|$ is the number of element in the nearest neighbor set.

Inspecting the right plot in Figure 9, a set of nearest neighbors for observation 26 is identified with dashed line and is denoted as:

$$N_k(p_{26}) = \{5, 25, 4, 3, 6\}$$

The $SBNpath_{26}$, ordering the observations in ascending order by distance to p_{26} , is:

$$SBNpath_{26} = \{25, 4, 5, 3, 6\}$$

The $SBNtrail_{26}$, identifying the shortest path for observations chaining distance, with solid line, is:

$$SBNtrail_{26} = \{(26, 25), (25, 4), (4, 3), (4, 5), (5, 6)\}$$

The cost function, returning the distance from pairs in $SBNtrail_{26}$, is:

$$c_{26} = \{1.7, 0.88, 1.08, 0.96, 0.97\}$$

The average chaining distance is then given by:

$$acDist_{26} = \left(\frac{2(5+1-1)}{5(5+1)} * 1.7 \right) + \left(\frac{2(5+1-2)}{5(5+1)} * 0.88 \right) + \dots = 6.07$$

and the connectivity-based outlier factor is then:

$$COF_5(p_{26}) = 1.52$$

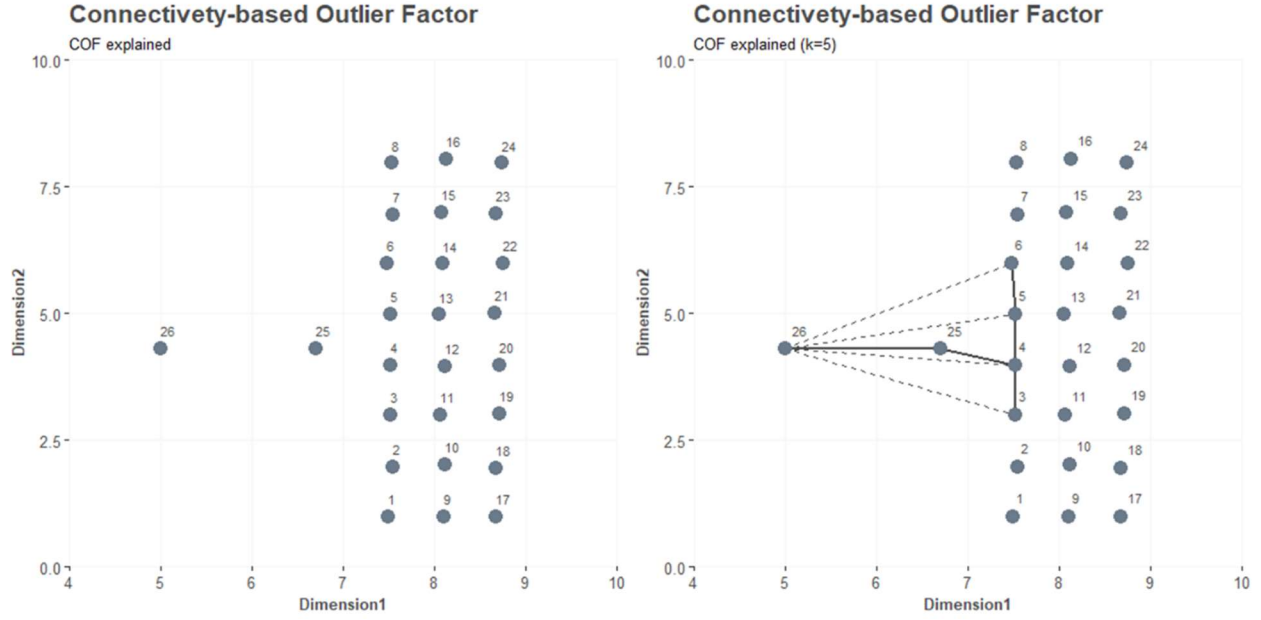


Figure 9: Dataset with observation 26 as outlier (left) and COF applied with $k=5$ (right)

Like INFLO and LOF, a COF of 1 indicates density – or isolativity – at the level of neighboring observations and COF greater than 1 indicates less dense neighborhood and outlierness. With the greatest COF, observation 26 ruled as the observation with the greatest outlierness. Again, as with LOF and INFLO, the disadvantage of COF is the subjectivity when choosing k .

For simplification, the LOF, COF and INFLO are described in one sentence:

- LOF identifies neighbors and express the ratio of distance between neighbors, and the neighboring observations of neighbors.
- COF identifies neighbors and express the ratio of distance between the most efficient chain for neighbors, and the most efficient chain for neighbors of neighbors.
- INFLO identifies reverse neighbors and express the ratio of distance between reverse neighbors, and the reverse neighboring observations of reverse neighbors.

For demonstration purposes of INFLO and COF, Figure 10 is created. The left plot shows few observations labelled with an outlierness ≥ 1.20 , being the observations with greatest deviating characteristics. Three observations are identified as inliers between cluster in bottom left and the two outlying observations

assigned with an INFLO of 1.49 and 1.26. Returning to the idea of the reverse neighborhood, the three inlying observations identifies the two outlying observations as a part of their reverse neighborhood, explaining the low INFLO score. For comparison, right plot has the very same three observations assigned with a relatively high COF score, indicating outlierness. Thus, INFLO and COF has different ways of defining outlierness. In this particular case, the incoherence for assigned outlierness for the two algorithms is also a matter of setting the k parameter right.

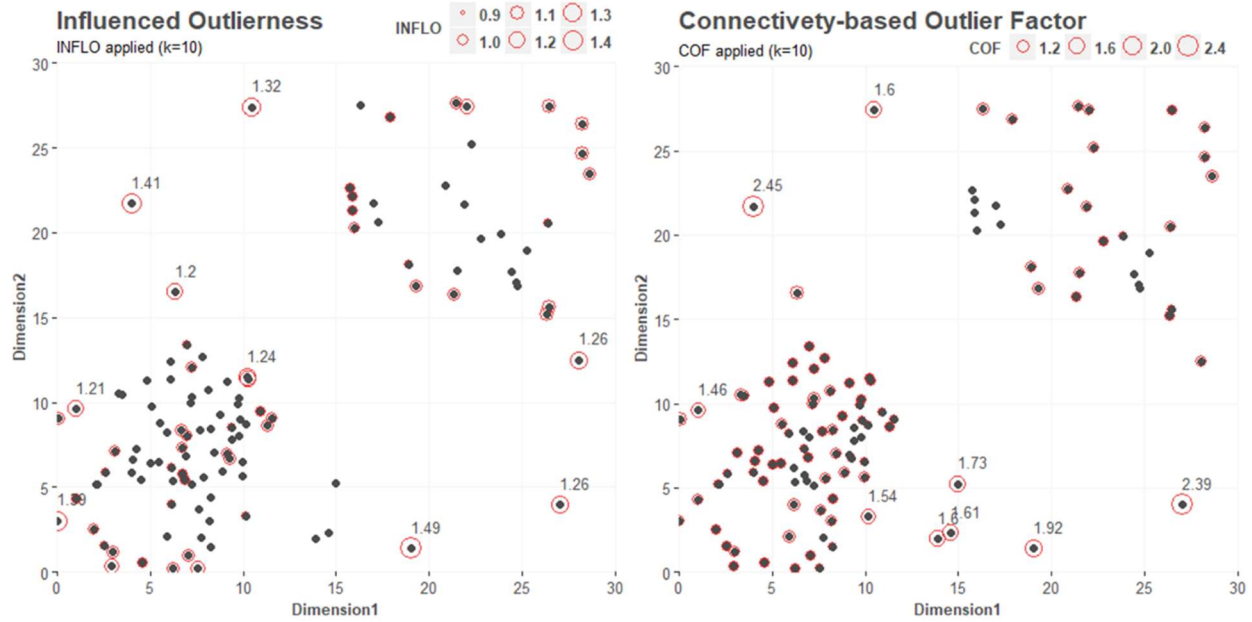


Figure 10: Comparison of INFLO (left) and COF (right) applied with $k=10$

3.2.7.4. Local Correlation Integral Method (LOCI)

Papadimitriou, Gibbons, & Faloutsos (2003) also adopted the density-based approach and criticized LOF for sensitivity when choosing k . It was argued that selecting a k larger than cluster size of which an observation belongs to will create artificial high outlier-score. A small k will on the other hand fail to identify small outlying clusters. Furthermore, the ‘magical cut-off point’ for the LOF-score to identify outliers was criticized. Papadimitriou, Gibbons, & Faloutsos (2003) proposed an algorithm, based on probabilistic reasoning, called Local Correlation Integral method (LOCI).

For probabilistic reasoning, a *sample* and *counting* neighborhood is introduced. For observation p_i , the sample neighborhood is given by $n(p_i, r)$, where r is a radius. The counting neighborhood is given by $n(p_i, \alpha r)$, where α is number between 0 and 1, determining the size of the sampling neighborhood,

relatively to the counting neighborhood. $\alpha = 0.5$ is recommended to have a sampling neighborhood large enough to contain samples.

For each identified observation p_n , within the sample neighborhood of observation p_i , the set of neighbors is expressed as $N(p_i, r) = \{p_n \in n(p_i, r) | \text{dist}(p_n, p_i) \leq r\}$. The count of neighbors in the set is expressed as $|N(p_i, r)|$. For each member of $N(p_i, r)$, a counting neighborhood is given by $n(p, \alpha r)$. The number of neighbors in $n(p, \alpha r)$ over the count of the sampling neighborhood is then given by:

$$\hat{n}(p_i, r, \alpha) = \frac{\sum_{p \in N(p_i, r)} n(p, \alpha r)}{|N(p_i, r)|}$$

The multi-granularity deviation factor, expressing the density of the counting neighborhood for observation p_i relatively to the counting neighborhood for observation p_n , is given by:

$$MDEF(p_i, \alpha, r) = 1 - \frac{n(p_i, \alpha, r)}{\hat{n}(p_i, r, \alpha)}$$

The counting neighborhood of observation p_i , always contains p_i . Thus, $\hat{n}(p_i, r, \alpha)$ is always > 0 . For the input parameter r , determining the size of both sampling and counting neighborhood, an r large enough to identify 20 members of $N(p_i, r)$ ($|N(p_i, r)| \geq 20$), was recommended as being enough to introduce statistical errors in $MDEF$ and $\sigma MDEF$.

To avoid the ‘magical cut-off point’, a normalized standard deviation of the counting neighborhood $\hat{n}(p_i, r, \alpha)$ was introduced, given by:

$$\sigma MDEF(p_i, \alpha, r) = \frac{\sigma \hat{n}(p_i, r, \alpha)}{\hat{n}(p_i, r, \alpha)}$$

If $MDEF$ is sufficiently large, an observation is labelled as an outlier given by:

$$MDEF(p_i, \alpha, r) > k_\sigma \sigma MDEF(p_i, \alpha, r)$$

where $k_\sigma = 3$ is recommended as less than 1% of the observations in a normal distribution are beyond the .999 confidence interval.

Inspecting the left plot in Figure 11, a sampling neighborhood $n(p_i, r)$ for observation p_i is given by $r = 2.25$. The counting neighborhood is given by $\alpha = 0.4$, or 40% of sampling neighborhood. Within the sample neighborhood of observation p_i , a set of neighbors are identified, $N(p_i, r) = \{p_i, p_1, p_2, p_3\}$. For

$n(p_1, \alpha r)$ and $n(p_2, \alpha r)$ 7 and 5 observations are identified, respectively. $n(p_i, \alpha r)$ and $n(p_3, \alpha r)$ both have 1 observation in their counting neighborhood, hence $\hat{n}(p_i, r, \alpha) = (1+7+5+1)/4 = 3.5$.

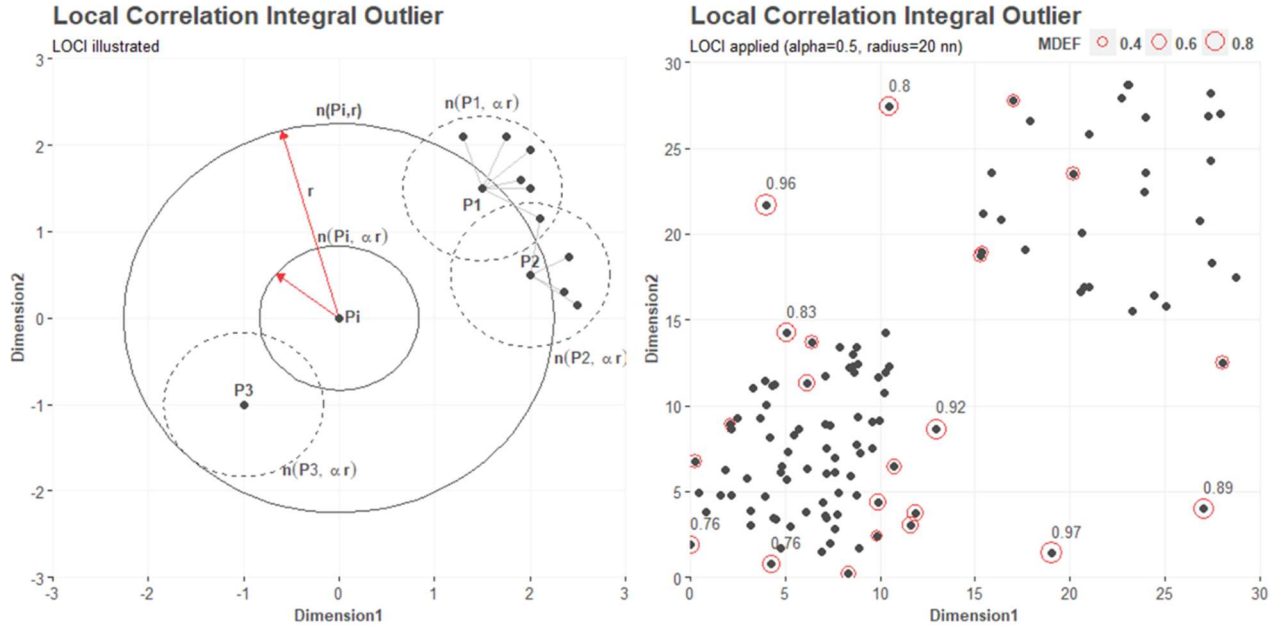


Figure 11: LOCI theory illustrated (left) and applied on synthetic data (right)

The multi-granularity deviation factor for observation p_i is then:

$$MDEF(p_i, \alpha, r) = 1 - \frac{n(p_i, \alpha, r)}{\hat{n}(p_i, r, \alpha)} = 1 - \frac{1}{3.5} = 0.71$$

Meaning that observation p_i has a less dense counting neighborhood than p_1, p_2 and p_3 . The standard deviation for $\hat{n}(p_i, r, \alpha)$ is given by:

$$\sigma \hat{n}(p_i, r, \alpha) = \sqrt{\frac{\left(\sum_{i=1}^{n=4} n(p, \alpha r) - \hat{n}(p_i, r, \alpha)\right)^2}{|N(p_i, r)| - 1}} = 3$$

And the normalized standard deviation is given by:

$$\sigma MDEF(p_i, \alpha, r) = \frac{\sigma \hat{n}(p_i, r, \alpha)}{\hat{n}(p_i, r, \alpha)} = \frac{3}{3.5} = 0.85$$

Thus, following the rule of a sufficiently larger MDEF with $k = 3$, observation p_i is not labelled as an outlier.

LOCI was introduced to decrease user subjectivity, by avoiding ‘magical cut-off point’ and k-NN parameter. Returning to figure 14, three observations in the right plot are identified with $MDEF \approx 1$, indicating neighboring observations in average have twice as dense neighborhood. However, following the suggested $k_\sigma = 3$, no observations are identified as outliers although some observation clearly have deviating characteristics, initiating a discussion for the value of k_σ . Furthermore, Papadimitriou, Gibbons, & Faloutsos (2003) do not provide any proof or argument for a radius based on 20 nearest neighbors, leading to a discussion of whether the value for radius should be given in distance, rather than a nearest neighbor parameter. A constant distance for each observation, possess a risk for outlying observations not to identify any neighbors in the sample neighborhood for comparison. On the other hand, a constant distance leaves the user with even more subjectivity, than a nearest neighbor parameter. Thus, LOCI is an incomplete theory, where MDEF seems to provide the best output for the user to judge outlierness.

3.2.7.5. Local Outlier Probability (LOOP)

As a response to LOCI, H-P. Kriegel, Kröger, Schubert, & Zimek (2009) proposed a different method that combines probability statistics and the local density-based element as introduced in LOF and its variants (i.e. COF, INFLO). H-P. Kriegel, Kröger, Schubert, & Zimek (2009) argued that a *probabilistic approach means to offer a natural tolerance to noise effects in the data* and that the output of LOF is sensitive to an inappropriate choice of k . For a slightly more robust method, offering an output between 0 and 1, a *probabilistic distance* is introduced.

Let $d(o, S)$ be the distance between observation o , of which outlierness is assigned to, and the set S , being observations in the nearest neighborhood. The number of observation in the nearest neighborhood is given by parameter k , being an input parameter. Assuming that o is in the center of nearest neighbor set S , the distances from o to $s \in S$ is approximately gaussian. A *standard distance* is then denoted as:

$$\sigma(o, S) = \sqrt{\frac{\sum_{s \in S} d(o, S)^2}{|S|}}$$

where $|S|$ is the number of elements in set S . The standard distance is slightly different from standard deviation, as it uses mean of 0 instead of the expected value $E[d(o, S)]$, based on the assumption of neighboring observations S being normally distributed around o . The *probabilistic set distance* of o to neighbors S is denoted as:

$$pdist(\lambda, o, S) = \lambda * \sigma(o, S)$$

where λ is an input parameter usually set to 1, 2 or 3, based on the three standard deviation rule ($\varphi \approx 68\%, 95\%, 99.7\%$). Thus, when $\lambda = 3$, 99.7% of the observations in S are within $pdist$. The *probabilistic local outlier factor* (PLOF) of observation o , returning the ratio of the density around observation o , relatively to the density for neighboring observations S , is defined as:

$$PLOF_{\lambda, S(o)} = \frac{pdist(\lambda, o, S(o))}{E_{s \in S(o)}[pdist(\lambda, s, S(s))]} - 1$$

where $E_{s \in S(o)}[pdist(\lambda, s, S(s))]$ is the expected value (mean) of $pdist$ for neighboring observations of observation o . PLOF has strong similarities to LOFs output, with its comparison of neighborhoods. To normalize PLOF, an aggregate value $nPLOF$ is defined as:

$$nPLOF = \lambda * \sqrt{E[(PLOF)^2]}$$

Thus, the $nPLOF$ is a kind of standard deviation for PLOF values.

With the means of an gaussian error function, to obtain a probability value between 0 and 1, the *Local Outlier Probability* (LOOP), indicating outlierness for observation o , is defined as:

$$LOOP_o = \max \left\{ 0, \text{erf} \left(\frac{PLOF_{\lambda, S(o)}}{nPLOF * \sqrt{2}} \right) \right\}$$

where erf is a gaussian error function.

3.2.8. High-dimensional approaches

For outlier detection suited for high-dimensional data, several methods have been proposed based on different methodologies. The methods only have an advantage in high-dimensional data in common, leaving the high-dimensional domain definition with ambiguity. High-dimensional data itself is also an ambiguous term. H-P. Kriegel, Schubert, & Zimek (2008) used the term for a minimum of 25 dimensions, Filzmoser, Maronna, & Werner (2006) used it for 50 dimensions and Lazarevic and Kumar (2005) does not mention the number of dimensions at all. Thus, high-dimensional approaches do not share a unified philosophy of an outlier as distance- and density-based approaches do.

3.2.8.1. Feature Bagging Outlier Detection (FBOD)

Feature bagging is a method widely used within the domain of predictive modelling to create a single *ensemble*, meaning a single model, by combining the result of other models. Lazarevic and Kumar (2005)

argued that utilizing diversity of different outlier algorithms by combining them will produce better results for identifying outliers.

Lazarevic and Kumar (2005) argued that not all features are relevant for detecting outliers and called for a method using a subspace of features to reduce the dimensions. Unlike standard bagging approach, it was proposed to use different features (variables) for each outlier detection algorithm, thus calling it *feature bagging*. The philosophy is to exclude redundant features and identify outliers as observations with deviating characteristics within the subspace, normally not detectable in high-dimensional space.

For outlier feature bagging, a combination of outlier detection algorithms in T rounds is proposed. For every round T , an algorithm is applied to a different set of features, denoted as $F_t = \{f_i, f_i \dots, f_i\}$. For each T , the number of features in F_t , denoted as N_f , is randomly chosen to a value between $D/2$ and $D - 1$, where D is the number of features in the original data. Each outlier detection algorithm assigns an outlier score to observations, denoted as AS_t , reflecting the outlierness for each observation. To unify AS_t scores for observations, two different methods are proposed; *breadth-first* and *cumulative sum*.

For each observation, the method called *breadth-first* standardize output of the outlier detection algorithms AS_t with indices, to a corresponding number of vectors denoted as Ind_t . The greatest value for Ind_t indicates outlierness and the smallest value indicates inlierness. Each observation is then assigned with the greatest Ind_t observed for the specific observation. For observation $n_1 = \{Ind_1, Ind_2, Ind_3\} = \{0.21, 0.95, 0.85\}$, 0.95 is returned as the *breadth-first* outlierness score.

For each observation, the method called *cumulative sum* sums the output of outlier detection algorithms AS_t in to a single variable AS_{FINAL} given by: $AS_{FINAL}(i) = \sum_{t=1}^T AS_t(i)$, returning the cumulative outlier score.

With feature bagging, Lazarevic and Kumar (2005) have successfully applied theory from a foreign domain to the discipline of outlier detection. The intuition of using multiple outlier detection algorithms to identify outliers is simple and provides robustness. Feature bagging has its strength in combining different algorithms in to a single output, allowing each algorithm to contribute. However, the theory has shortcomings. If only algorithms with continuous output is used as input algorithms, the algorithms provide output in different units and variance; k-NN Graph returns an in-degree, standard k-NN returns a LOF value etc. Thus, to judge outlierness, it is recommended to standardize the AS_t output as Ind_t and average or sum it.

3.2.8.2. Angle-Based Outlier Detection (ABOD)

H-P. Kriegel, Schubert, & Zimek (2008) argued that concepts as proximity, distance and nearest neighbors become less meaningful with increasing dimensionality of datasets. This is described as the curse of dimensionality leading to the problem called *Data-Snooping Bias* where every point is an outlier, given enough subspaces (Zimek, Schubert, & Kriegel, 2012). Thus, the concept of neighborhood becomes meaningless in high-dimensional data.

To overcome this shortcoming for distance- and density-based approaches, an algorithm based on angle towards other observations was proposed. The logic is that an inlier will be centered in the data, having strong directional variance in the angles towards other observations, while an outlier will have small directional variance in the angle towards other observations. For observation $\vec{p}$ in a dataset, given observation $\vec{b}$ and $\vec{c}$ for comparison, the variance of the angles between the two vectors $\vec{b}$ and $\vec{c}$, is given by:

$$ABOD(\vec{p}) = VAR_{\vec{b}, \vec{c} \in D} \left(\frac{\langle \overline{p\vec{b}}, \overline{p\vec{c}} \rangle}{\|\overline{p\vec{b}}\|^2 * \|\overline{p\vec{c}}\|^2} \right)$$

where $\overline{p\vec{b}}$ is difference for vector $\vec{p}$ and $\vec{b} = (\vec{p} - \vec{b})$. A small ABOD is equal to small variance, meaning it is an outlier while a great ABOD is an inlier.

Inspecting Figure 12, which was created for demonstration purpose, the left plot shows observation p as an outlier with a narrow angle towards other observations. With an ABOD of .004, it has small variance in the angles. The right plot shows observation p with a wide angle towards other observations. With an ABOD of 0.77 it has strong variance in the angles, hence being an inlier.

The advantage of ABOD is the absence of input parameters and its continuous output, allowing the user to rank the outlierness for each observation. Although it is stated that computational properties are not discussed, it should be mentioned that ABOD is practically impossible to compute with more than 5.000 observations and 10 dimensions. A version with a k parameter for k nearest neighbors is therefore also suggested by Kriegel et. al. Moreover, it appears that ABOD is only applicable in high-dimensional data. For data given in Figure 4, ABOD will fail to identify the observations with deviating characteristics in-between clusters of data.

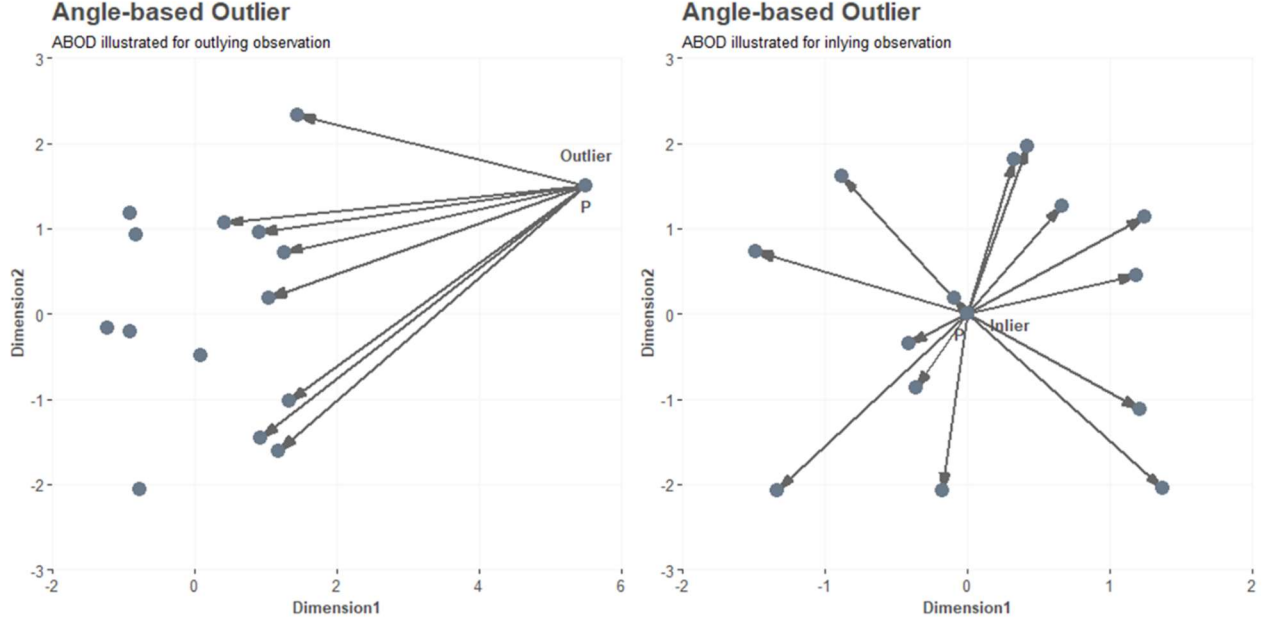


Figure 12: ABOD illustrated for outlying (left) and inlying observation (right)

3.2.8.3. Subspace Outlier Degree (SOD)

Again, referring to curse of dimensionality, H.-P. Kriegel, Kröger, Schubert, & Zimek (2009) introduced Subspace Outlier Degree (SOD) as a tool to identify outlierness for observations. It was argued that variables with uniform or normal distributions do not contain outliers. Instead, the SOD focus on the variables with deviating distributions, being the so-called *subspace*.

The proposed algorithm defines the subspace as S . The variance for subspace S is the average squared distance from observations p to the respective μ , given by:

$$VAR^S = \frac{\sum_{p \in S} dist(p, \mu)^2}{Card(S)}$$

where $Card(S)$ is the cardinality of S : the number of dimensions in the subspace. For outlier reference, a set of observations is denoted as $R(p)$, for observation p to be compared with. This set of observations can take any number of observations (max $n - 1$) and is an input parameter.

For $R(p)$ a vector defining the *actual subspace* (i.e. the dimensions/variables) is denoted as $v^{R(p)}$. The variance of $R(p)$ is denoted as $VAR^{R(p)}$ and for each dimension in $v^{R(p)}$, the average variance is denoted as: $\frac{VAR^{R(p)}}{d}$.

For choosing dimensions to the *subspace hyperplane* $v^{R(p)}$, for observation p to be compared with $R(p)$, dimensions are included if $VAR_i^{R(p)} < \alpha \frac{VAR^{R(p)}}{d}$, where α is an input parameter. Thus, to compare observation p with reference set $R(p)$, the variance for dimensions d in $v^{R(p)}$ is desired low.

The *subspace hyperplane* $H(R(p))$ of $R(p)$ is defined by the mean value $R(p)$, denoted as $\mu^{R(p)}$, and the subspace defining vector $v^{R(p)}$, leading to the notation $H(R(p)) = (\mu^{R(p)}, v^{R(p)})$. For the reference set $R(p)$, mean is given by $\mu^{R(p)}$, while value for observation p is given by p_i . The deviation of observation p to reference set $R(p)$ on subspace hyperplane is given with Euclidean distance:

$$dist(p, R(p)) = \sqrt{\sum_{i=1}^d v_i^{R(p)} * (p_i - \mu_i^{R(p)})^2}$$

The final output of the algorithm, the SOD score, is denoted as:

$$SOD_{R(p)}(p) = \frac{dist(p, R(p))}{\|v^{R(p)}\|}$$

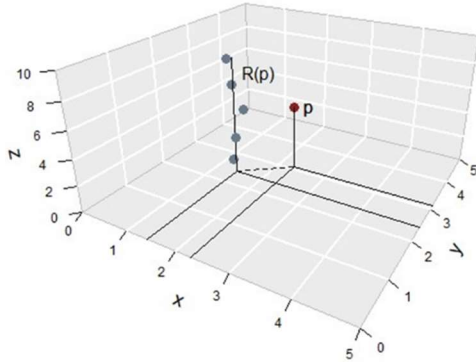
where $\|v^{R(p)}\|$ denotes the number of dimensions chosen when $VAR_i^{R(p)} < \alpha \frac{VAR^{R(p)}}{d}$.

The left plot in Figure 13 shows a 3-dimensional data space where observation p is compared to reference set $R(p)$. The reference set has a variance, denoted as $VAR_i^{R(p)}$, for dimension x , y and z of 0.003, 0.003 and 8, respectively. With an average variance, $\frac{1}{d} * VAR^{R(p)}$, of 2.66 and $\alpha = 0.8$, as recommended by Kriegel et. al, dimensions x and y are retained for evaluation of outlieriness for observation p , relative to reference set $R(p)$.

The right plot in Figure 13 shows the subspace hyperplane, consisting of dimension x and y , where the reference set $R(p)$ has minimal variance. The Euclidean distance between observation p and reference set $R(p)$, denoted as $dist(p, R(p))$, is 1.04. SOD score is then: $SOD_{R(p)}(p) = \frac{1.04}{2} = 0.52$, returning the average deviation per dimension from observation p to reference set $R(p)$, where $R(p)$ has minimal variance.

Subspace Outlier Degree

Subspace for observation p



Subspace Outlier Degree

Subspace hyperplane for observation p

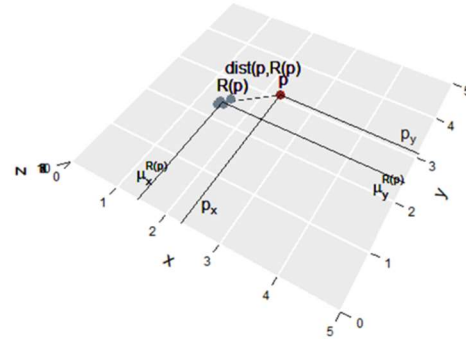


Figure 13: Subspace Outlier Degree illustrated with subspace (left) and subspace hyperplane (right)

3.3. CLUSTERING

3.3.1. Clustering objectives

On geography and spatial proximity, Tobler (1970) stated that ‘everything is related to everything else, but near things are more related than distant things’ with the first law of geography. Introduced by Driver and Kroeber (1932), clustering was first used within the field anthropology and has since been adopted by other fields of work, while the literature is still undergoing development today. Although no exact definition of clustering exists, it is widely expressed that clustering is a tool used to identify similarities among observations and summarize dimensions in a multidimensional space.

“The objective of the clustering methods is to discover significant groups present in a data set. In general, they should search for clusters whose members are close to each other (in other words have a high degree of similarity) and well separated”

”

- Halkidi, Batistakis, & Vazirgiannis (2001)

“Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in the other

groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering ”

– Tan, Steinbach, Karpatne, & Kumar (2005)

“ *Clustering is another application of our fundamental notion of similarity. The basic idea is that we want to find groups of objects (consumers, businesses, whiskeys, etc.), where the objects within groups are similar, but the objects in different groups are not so similar* ”

– Provost & Fawcett (2013)

Thus, clustering combines observations in to groups (i.e., clusters) such that:

- Each cluster is homogenous or compact with respect to certain characteristics. That is, observations in each cluster are similar to other observations in same cluster. The objective is operationalized as minimizing intra-cluster distance.
- Each cluster is different from other clusters with respect to the same characteristics. That is, observations of one cluster should be different from observations in different clusters. The objective is operationalized as maximizing inter-cluster distance.

3.3.2. Clustering evaluation

“ *The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage* ”

- Jain & Dubes (1988)

To evaluate the criteria of minimizing intra-cluster distance and maximizing inter-cluster distance, several metrics have been identified. Metrics are divided in *external* and *internal* (Tan et al., 2005). External evaluation metrics cover the extent to which a clustering solution fits an external classification and hence have strong similarities with supervised learning. Internal evaluation metrics cover the fitness of a clustering solution without any external information. Internal evaluation metrics are divided in to *relative* and *absolute* metrics. Relative evaluation metrics covers comparison of different clustering solutions (i.e.

different applied algorithms) and usually provide an output between 0 and 1, while absolute metrics provides an absolute figure.

With respect to the stated objective for the thesis at hand, external evaluation metrics are excluded for usage in analysis.

3.3.2.1. Internal absolute metrics

To evaluate the criterion of minimizing intra-cluster distance, *Sum of Squares Within (SSW)* is useful. *SSW* sums the variance withheld in the clusters, as a measure of information and is defined as:

$$SSW = \sum_{g=1}^G \sum_{i \in C_g} \sum_{k=1}^p (x_{ik} - \bar{x}_{gk})^2$$

where C_g is notation of cluster g , $i \in C_g$ is observation i as a member of cluster C_g and k is the k th dimension. Thus, *SSW* squares the distance from the i th observation of the k th dimension, denoted as x_{ik} , to the center of cluster g of which observation x_{ik} is a member of, for the respective k th dimension, denoted as $\bar{x}_{gk}$. A low *SSW* is desired for low variance and dense clusters.

3.3.2.2. Internal relative metrics

Davies and Bouldin (1979) argued that previous developed clustering evaluation metrics depended on the determination of relatively large changes in an index, rather than its minimization or maximization. To overcome this issue, the Davies-Bouldin index was introduced. The Davies-Bouldin index compares the intra-cluster distance and inter-cluster distance and is defined by following:

$$DaviesBouldin = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

where n is the number of clusters, $\sigma_i + \sigma_j$ is the standard deviation of observations in cluster i and j , to their respective centroid c_x and $d(c_i, c_j)$ is the distance between cluster c_i and c_j . For each pair of clusters for comparison, being c_i and c_j , the pairs of clusters maximizing Davies-Bouldin is summed.

Another relative index is the Dunn index (Dunn, 1974). The Dunn index compares the minimum inter-distance for all pairs of clusters with the greatest intra-distance for all given clusters and is given by following formula:

$$Dunn = \frac{\min_{i \neq j} \{d_c(C_i, C_j)\}}{\max_{1 \leq l < k} \{\Delta(C_l)\}}$$

where $\min_{i \neq j} \{d_c(C_i, C_j)\}$ is the minimum inter-cluster distance observed between all pairs of clusters and $\max_{1 \leq l < k} \{\Delta(C_l)\}$ is the greatest intra-cluster distance observed in all clusters, given by the diameter $\Delta(C_i)$. The greater the Dunn index, the greater clustering results are achieved with respect to inter- and intra-cluster distance.

The Dunn index has a shortcoming, as it is bound to identify the worst structure, between all pairs of clusters for inter-cluster distance, given $\min_{i \neq j} \{d_c(C_i, C_j)\}$. Thus, it is strongly biased to point out a ‘local’ cluster solution and return a non-satisfactory index, leaving out the global comparison. Furthermore, noise or a single complex shaped cluster will increase the diameter ($\Delta(C_i)$), worsening the result with respect to intra-cluster distance. Dunn is therefore not a robust evaluation metric.

In addition to the absolute measure SSW , several authors have suggested the R^2 measure (Aggarwal & Reddy, 2013; Halkidi et al., 2001; Tan et al., 2005), as an expression of total explained variance. R^2 is defined as:

$$R^2 = \frac{SSB}{SST}$$

where $SST = Total\ Sum\ of\ Squares$:

$$SST = \sum_{i=1}^n (x_i - \bar{x})^2$$

and $SSB = Sum\ of\ Squares\ Between$:

$$SSB = \sum_{i=1}^n |c_i| (x_i - \bar{x})^2$$

where $|c_i|$ is the number of members for cluster c_i .

Thus, SST is a measure of total variance in data, returning the squared distance from observations to the center of gravity while distance between clusters is defined as the distance to the center of data. R^2 therefore differ in its definition of inter-cluster distance, when compared to Dunn and Davies Bouldin, as inter-cluster distance simply is defined as the distance to the center of gravity for each cluster.

With respect to minimizing intra-cluster distance and maximizing inter-cluster distance, a higher R^2 is desired, as minimizing SSW is equivalent to maximizing SSB.

3.3.2.3. Internal relative metrics - cluster belongingness

For evaluating clustering results with respect to the identified objectives, metrics dealing with observations belongingness to a certain cluster are also useful. The Silhouette coefficient (Rousseeuw, 1987) is a coefficient indicating belongingness for observations to an assigned cluster and is defined by following:

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

where $b(i)$ is the minimum distance of observation i to all observations in nearest clusters of which i is not a member and $a(i)$ is the average distance of observation i to all observations in same cluster.

The coefficient ranges from -1 to +1, where +1 indicates strong similarities with other observations in same cluster and -1 indicates no similarities with other observations in same cluster. Naturally, for $S(i)$ to be greater than 0, it is required that the average distance for i to other observations in same cluster is lower than minimum distance of observation i to all observations in nearest cluster of which i is not a member, such that $(a(i) > b(i))$. An observation with coefficient < 0 , is perceived as belonging to a neighboring cluster. For general evaluation of a clustering results, the average coefficient of all observations is used.

Relying on a observations-to-nearest-cluster-observations comparison for each observation i , the Silhouette coefficient differs from the local cluster-to-cluster method Dunn. Thus, the Silhouette coefficient is more robust than the Dunn index.

Another metric dealing with cluster belongingness is the cluster correlation developed by Tan, Steinbach, Karpatne, & Kumar (2005). To determine to which extent clusters are well-separated, correlation between a distance and incidence matrix is suggested.

Let the incidence matrix be a matrix of $n \times n$ with each element returning 1 if two observations are in same cluster and let the distance matrix be a matrix of corresponding size ($n \times n$). For demonstration purpose, random incidence and distance matrix when $n = 5$ is given:

$$IncMatrix = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \end{bmatrix}, \quad DistMatrix = \begin{bmatrix} 0 & 0.3 & 5 & 3 & 0.4 \\ 0.3 & 0 & 4 & 4.5 & 0.2 \\ 5 & 4 & 0 & 0.4 & 2.1 \\ 3 & 4.5 & 0.4 & 0 & 3 \\ 0.4 & 0.2 & 2.1 & 3 & 0 \end{bmatrix}$$

The *IncMatrix* reveals following pairs of observations belonging to same clusters:

$$Pairs = \{ (1,2), (1,5), (2,5), (3,4) \}$$

Leading to two clusters:

$$C_1 = \{1,2,5\}, \quad C_2 = \{3,4\}$$

By vectorizing the two matrices, following vectors are obtained:

$$IncVector = \{1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1\}$$

$$DistVector = \{0, 0.3, 5, 3, 0.4, 0.3, 0, 4, 4.5, 0.2, 5, 4, 0, 0.4, 2.1, 3, 4.5, 0.4, 0, 3, 0.4, 0.2, 2.1, 3, 0\}$$

With a correlation of:

$$cor(IncVector, DistVector) = -0.92$$

Meaning that observations paired with 1 in *IncMatrix*, correlates negatively with distance between the very same observations. That is, observations in same cluster are closer than observations in other clusters.

3.3.2.4. Introducing: Cluster Index

For the identified tools for cluster evaluation, Silhouette coefficient is identified as having great properties with respect to robustness, being a method for each observation i , while Davies-Bouldin relies on a cluster-to-cluster comparison. Moreover, for Davies-Bouldin the following are identified:

- Davies-Bouldin assigns the greatest value to each cluster, when compared with other clusters. One cluster can have great distance to other clusters, and still have a high Davies-Bouldin assigned if just another cluster is close. Thus, Davies-Bouldin does not take the distance to other clusters in to account, but solely base the index on the worst possible cluster for comparison.
- Davies-Bouldin assigns every cluster with equal weight, when compared with other clusters, regardless of cluster size. A small cluster with great intra-cluster distance close to cluster of which

Davies-Bouldin is assigned to, possess a risk of returning a large Davies-Bouldin. Thus, Davies-Bouldin does not take the proportion of observations in neighboring clusters in to account, but risk comparing with clusters accounting for an insignificant proportion of the total observations.

To overcome the issue of missing robustness, a simple but robust method is proposed by Tan, Steinbach, Karpatne, & Kumar (2005), suggesting using the two metrics; *cohesion* and *separation*. Cluster cohesion is defined as the sum of all links (distances) within a cluster, while clustering separation is defined as the sum of all links to observations in other clusters.

Cluster cohesion is denoted as:

$$Cohesion(c_i) = \sum_{\substack{x \in c_i \\ y \in c_i}} dist(x, y)$$

where c_i is the cluster of which distance between observations is measured.

Cluster separation is denoted as:

$$Seperation(c_i, c_j) = \sum_{\substack{x \in c_i \\ y \in c_j}} dist(x, y)$$

where c_i is the cluster of which observations distance to observations in cluster c_j is measured.

The advantage of cohesion and separation is the weighting of every observation, making it more robust than Davies-Bouldin and Dunn relying on a cluster to cluster solution.

For further development and utilization of the theory, average cluster cohesion and cluster separation is introduced:

$$AvgSeperation = \frac{\sum_{\substack{x \in c_i \\ y \in c_j}} dist(x, y)}{|c_i| * |c_j|} = \frac{Seperation(c_i, c_j)}{|c_i| * |c_j|}$$

where $|c_i|$ is the number of members for cluster c_i and $|c_j|$ is the number of members for cluster c_j .

Average cluster cohesion is denoted as:

$$AvgCohesion = \frac{\sum_{\substack{x \in c_i \\ y \in c_i}} dist(x, y)}{|c_i| * (|c_i| - 1)} = \frac{Cohesion(c_i)}{|c_i| * (|c_i| - 1)}$$

where $|c_i|$ is the number of members for cluster c_i .

Returning to clustering objectives, the ratio between intra-cluster and inter-cluster distance is given by:

$$\text{Cluster index} = \frac{\text{AvgCohesion}}{\text{AvgSeperation}}$$

Thus, the cluster index is an index returning the average distance between observations in same cluster, over the distance for reaching observations in other clusters, of which observations do not belong to. Naturally, the index ranges from 0 to 1, with 1 being a poor solution and 0 being the optimal solution.

For demonstration purposes of clustering evaluation methods, Figure 14 is created. Plot one and plot three consist of clearly divided clusters. Plot three has a cluster with a convex shape and two gaussian clusters, while plot one consists of two gaussian clusters. Data for plot two and four consist of three and two gaussian clusters, respectively. Both plots have a low inter-cluster distance, and both have k-means applied with $k=3$, explaining the split of the two gaussian clusters - into three clusters – for plot four.

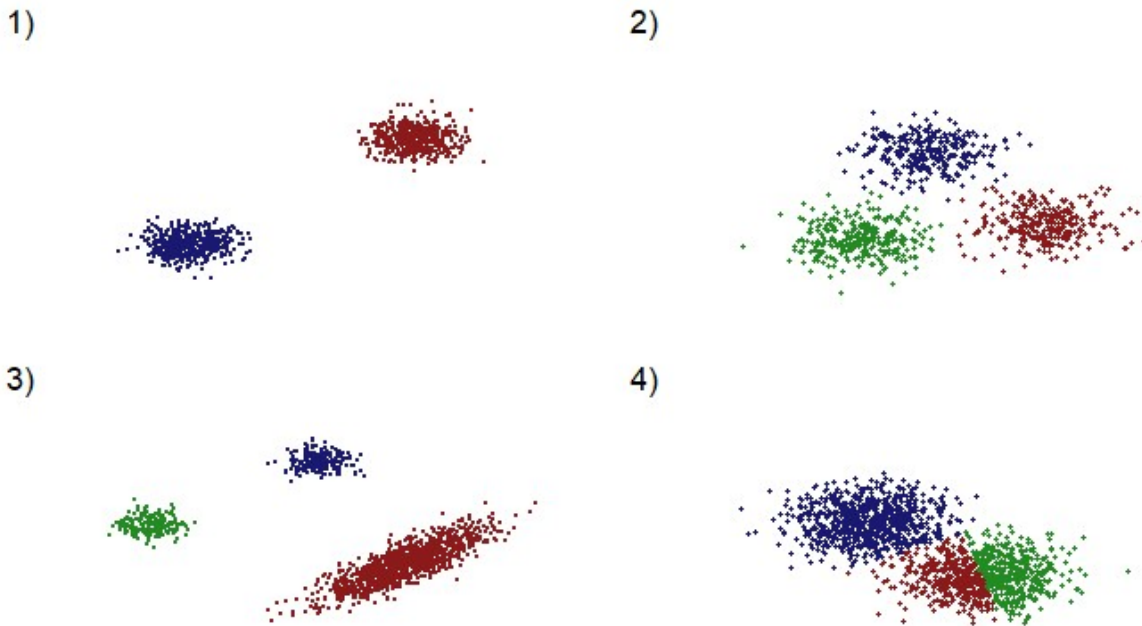


Figure 14: Clustering for synthetic data

Inspecting the corresponding cluster evaluation metrics in Table 2, the weakness of Dunn is exposed. Both plot one and plot three shows a great inter-cluster distance, but with a Dunn of 0.31, the cluster solution for plot three is evaluated significant worse than cluster solution for plot one (1.34). This has to do with

the beforementioned weakness of Dunn (i.e. taking the diameter of clusters in to account), meaning the convex shape for one of three clusters in plot three distorts the Dunn index to an extent where Dunn evaluates cluster results closer to plot two and four, than to plot one. Davies-Bouldin has greater properties to deal with the convex shape of clusters and assigns plot three with a Davies-Bouldin of 0.33, worse than plot one with a Davies-Bouldin of 0.20, but nowhere as close to plot two and plot four, with a Davies-Bouldin of 0.49 and 1.18, respectively. Thus, Davies-Bouldin recognize that plot three has worse intra-cluster distance than cluster one but does not assign a value indicating significant worse results.

Plot	Davies-Bouldin	Dunn	Silhouette	Cluster correlation	Cluster index	R ²
1	0.20	1.34	0.87	-0.98	0.12	0.96
2	0.49	0.05	0.66	-0.84	0.29	0.84
3	0.33	0.31	0.74	-0.91	0.26	0.84
4	1.18	0.00	0.41	-0.64	0.40	0.75

Table 2: Cluster evaluation metrics for synthetic data

For cluster belongingness, Silhouette index returns a Silhouette of 0.74 for plot three, indicating a worse cluster solution than for plot one. Again, the solution is evaluated closer to the cluster solution for plot two, although plot two has less inter-cluster distance. Returning to the explanation of Silhouette index, the Silhouette index identifies the minimum distance to an observation in another cluster and divide it with the average distance to observations in own cluster. Having in mind that intra-cluster distance is worse for plot three (i.e. long convex shape), than for plot two, it seems reasonable that the cluster evaluation of plot three is closer to cluster evaluation for plot two, than for plot one, as with Davies-Bouldin.

The cluster correlation of -0.98 for plot one indicates a somewhere near perfect cluster solution. For plot three, the correlation of -0.91, place the clustering solution in-between cluster solution for plot one and plot two (-0.84), while the cluster solution for plot four has a relatively poor correlation of -0.64. For plot four, the cluster correlation with a solution of two clusters instead of three, is -0.76 and Silhouette coefficient is 0.59 (not in table), indicating that a clustering solution with just two clusters for plot four is preferred.

The cluster index ranks the clustering solution of plot three, significant closer to clustering solution for plot two, than for plot one. With a cluster index of 0.26 for plot three, it is revealed that 26% of the average distance to observations in other clusters, is equal to the average distance to observations in same

clusters. Again, having in mind that intra-cluster distance is worse for plot three, than for plot two, it seems reasonable that the cluster index is not as great as for plot one. On the other hand, the clustering solution for plot two has a cluster index of 0.29, revealing that almost a third of the average distance to observations in other clusters is equal to the average distance for observations in own cluster. With less intra-cluster distance for plot two, when compared to plot three, the cause of the relatively poor cluster index of 0.29 is simply the denominator, being the average distance to observations in other clusters, reflecting the inter-cluster distance.

The R^2 evaluates the clustering solution of plot one somewhere near perfect with an R^2 of 0.96. As the only evaluation metric, the clustering solution for plot three and plot two are evaluated equal, while the clustering solution for plot four is evaluated worst. Returning to the explanation of R^2 , it is apparent that inter-cluster distance is defined as distance to the center of data, explaining why the poor inter-cluster distance for plot two does not influence the solution greatly.

Summing up the cluster evaluation tools, all six evaluation tools evaluate clustering solution for plot four worst, while clustering solutions for first plot is evaluated as being best with respect to intra- and inter-cluster distance. Davies-Bouldin, Silhouette and cluster index evaluates the clustering solution of plot three closest to the cluster solution of plot two, while correlation evaluates clustering solution of plot three as close to clustering solution for plot one as to plot two. Dunn evaluates the clustering solution of plot three closer to both clustering solution for plot two and plot four, than to plot one. R^2 evaluates the clustering solution of plot three and plot two equal.

3.3.3. Cluster tendency

To evaluate and ensure diversity in chosen datasets, it is necessary measure to which extent a dataset contains clusters. That is, to which extent a dataset is uniform distributed or partitioned in clusters. To measure cluster tendency, the numerical indicator Hopkins statistic (Hopkins, 1954) is applicable.

Let $p = \{p_1, \dots, p_n\}$ be a sample of n size from dataset D and let $q = \{q_1, \dots, q_n\}$ be an artificial created uniform distribution of n size within space in dataset D . For each point p_i , find the nearest neighbor in set p and denote distance as $x_i = \min_{p_i \in D} \{dist(p_i, p_i)\}$. For each point q_i , find the neighbor in set p and denote distance as $y_i = \min_{p_i \in D} \{dist(q_i, p_i)\}$. The Hopkins statistic is then defined as:

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$

Being a test for uniform distribution in data, the Hopkins statistics returns an index $\{0,1\}$, where 0.5 characterize random uniformly distributed data, 1 characterize data partitioned in clusters and 0 characterize random data.

For demonstration purpose, Figure 15 is created. A uniform distribution is given in blue, being $p = \{p_1, \dots, p_n\}$, with another uniform distribution in black, being $q = \{q_1, \dots, q_n\}$. The right plot has two gaussian clusters in blue and a uniform distribution in black, with same notation of p and q .

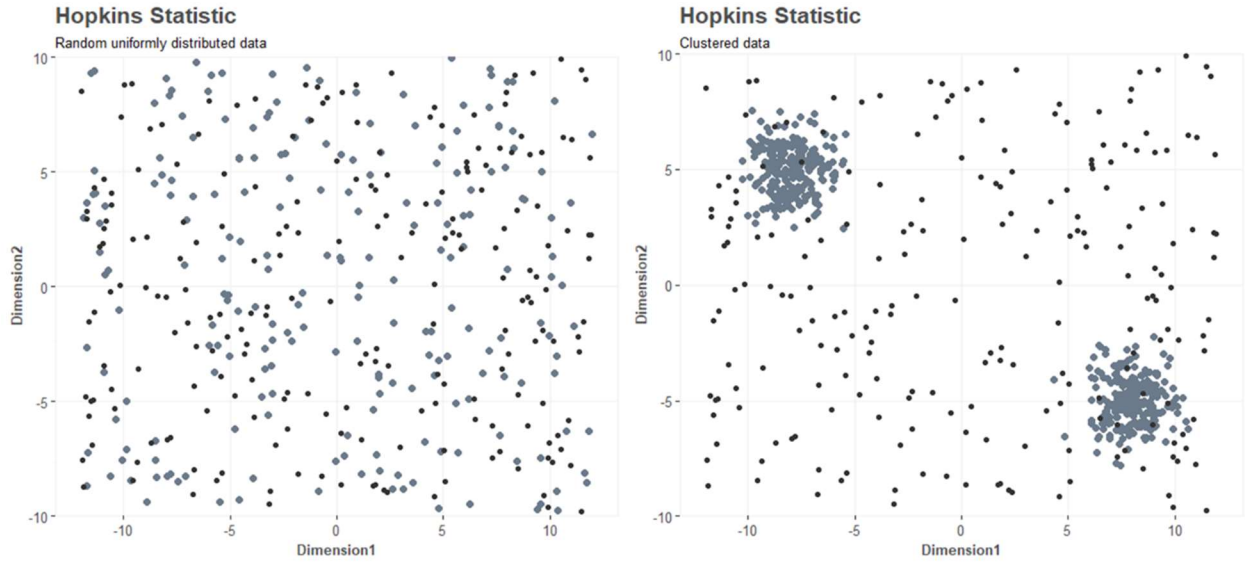


Figure 15: Uniform distribution in blue (left) and two gaussian clusters in blue (right)

For the right plot, the accumulated sum of distance from q_i to nearest neighbor in set p is:

$$\sum_{i=1}^n y_i = 240.3$$

The accumulated sum of distance from p_i to nearest neighbor in set p is:

$$\sum_{i=1}^n x_i = 16.7$$

Leading to a Hopkins statistic of:

$$H = \frac{240.3}{16.7 + 240.3} = 0.93$$

For the left plot, the corresponding Hopkins Statistic is 0.47, meaning the accumulated sum of distance from q_i to nearest neighbor in set p , is slightly smaller than accumulated sum of distance from p_i to

nearest neighbor in set p . Thus, for the right plot, hypothesis of uniform distributed data is rejected and dataset is considered to have strong cluster tendency, contrary to the left plot.

3.3.4. Clustering algorithms

As with outlier detection, clustering has been developed from different approaches with the most notable ones being hierarchical clustering, centroid-based clustering and density-based clustering. With respect to the objective and research question, the density-based clustering tool DBSCAN is left out as it is a clustering tool with built-in exclusion of noisy observations.

3.3.4.1. Hierarchical clustering

Hierarchical clustering is a clustering method working with a bottom-up approach, meaning no random seeds are used to shape clusters, leaving out the stochastic element. For identifying similarities among observations, hierarchical clustering takes an input parameter of a distance metric, with Euclidean distance being the most common. Using a distance matrix based on the chosen distance metric, points are clustered, starting from the two closest observations. To calculate the distance from clustered observations, to non-clustered observations, an input parameter of linkage method is selected. With a linkage method, a 'hierarchy' is built of clusters going from most granular to eventually one cluster being least granular. The algorithm is described with syntax in Table 3.

Hierarchical clustering algorithm
1. Select distance metric: $distMetric = \{Euclidean, Manhattan, Minkowski, Mahalanobis\}$
2. Select linkage method: $linkMethod = \{Centroid, Single, Complete, Average, Ward's\}$
3. Create $distMatrix$ for dataset based on $distMetric$
4. Repeat:
a. With $distMatrix$, merge two closest observations in to cluster
b. Update $distMatrix$ and calculate distance from cluster based on $linkMethod$
5. Until no more clusters exist and the hierarchy is complete

Table 3: Hierarchical clustering algortihm syntax

3.3.4.2. K-means clustering

K-means clustering is a clustering method working with a stochastic element. For identifying similarities among observations, k-means takes an input parameter of a number of seeds/centroids to be placed randomly. As with hierarchical clustering, a distance matrix based on the selected distance metric is created, and distance from centroids to observations is calculated. For x number of centroids,

observations are assigned as belonging to the closest centroid, based on distance. The centroids are then relocated to the center of the belonging observations, and the process repeats until the convergence criterion is achieved. K-means use arithmetic mean to relocate centroids. K-means algorithm is described with syntax in Table 4.

K-means algorithm
1. Select distance metric: <i>distMetric</i> = {<i>Euclidean</i>, <i>Manhattan</i>, <i>Minkowski</i>, <i>Mahalanobis</i>} 2. Select centroid method: <i>centroid</i> = {<i>arithmetic mean</i>} 3. Select number of centroids: N_c 4. Randomly place centroids within borders of data space 5. Repeat: a. Calculate distance from observations to <i>centroid</i> b. Assign each observation to nearest <i>centroid</i> c. Based on assigned observations, relocate <i>centroid</i> 6. Until convergence criteria of no further relocation for <i>centroid</i> is met 7. Label the assigned observations to clusters

Table 4: K-means algorithm syntax

3.3.4.3. Discussion of clustering algorithms

For optimal clustering results, hierarchical and k-means algorithm have advantages in different domains of data. Hierarchical clustering has an advantage in usability and interpretation. With different linkage methods such as single linkage, hierarchical clustering has an advantage when dealing with complex shapes. The greedy single linkage namely connects to closest observation, making it easy to form non-gaussian clusters. Moreover, when working with the dendrogram, hierarchical clustering also allows the user to judge similarities among observations on different levels (i.e., porpoise and dolphins are very similar but is a ‘subcluster’ to a whale mammals cluster). The dendrogram is useful within the domain of genetics, species and data where observations are labelled with a ‘title’ the user can relate to. The dendrogram has however no interpretation power in a customer database with million of records. As an disadvantage, hierarchical clustering is a static tool with constant outcome, working with a bottom-up approach, having difficulties identifying clusters of different densities.

K-means clustering works with a stochastic element where users select a number of centroids, consequently being the number of clusters. The greatest drawback of k-means is the sensitivity to the initial settings. With the stochastic element, k-means produces different results for each run and furthermore only returns results for the selected number of centroids. For the user to find the global

optimum, the user has no other options but to loop over different number of centroids many times. Results are therefore difficult to reproduce. Moreover, k-means is sensitive to outliers and has difficulties working with non-gaussian clusters and clusters of different sizes. As an advantage, centroid-based clustering is easy to compute and is more flexible in creating clusters.

4. METHODOLOGY FOR APPLYING THEORY

4.1. SELECTION OF DATASETS

For selection of datasets, a set of criteria are put in place:

1. Datasets must be diverse in composition of attributes and attribute characteristics

To evaluate algorithms efficiency when dealing with different attributes, it is striven for diversity for mix of continuous and categorical (discrete) attributes. Diversity in attribute characteristics, such as distribution properties, is emphasized to avoid performance bias for one algorithm.

2. Datasets must be different from datasets used in literature

To avoid confirmation bias for a single proposed algorithm, the datasets must be different from the datasets used in the identified literature.

3. Datasets must be diverse in size

To evaluate algorithms efficiency when dealing with data in different sizes, it is striven for a diverse mix of datasets with few and many dimensions.

For application of algorithms, the datasets have been treated with removal of irrelevant attributes such as ID (unique identifier for rows) and attributes with missing values. Any attribute with text values is converted to a dummy variable, such that $Species = \{setosa, versicolor, virginica\} \rightarrow \{1,2,3\}$

To ensure diversity, datasets are inspected from different perspectives:

- *Attribute composition* is the composition of continuous and categorical attributes.
- *Central tendency* is the tendency of values to center around the center of gravity. It is characterized by absolute average skewness (abbreviation: Skew.) and average kurtosis, for attributes in dataset. A great skewness and kurtosis indicates low central tendency.
- *Outliers* is the number of outliers according to Tukey's boxplot, with $k = 3$.
- *Cluster tendency* is tendency of clusters in dataset, measured by previously explained Hopkins statistic.

All datasets are from UCI Machine Learning repository. Similarity matrices for datasets are available in annex.

With less than 20 dimensions, six datasets are categorized as low-dimensional (Table 5). With only 1 and 2 categorical attributes, *PimaIndians*, *Winequality* and *ForestFires* predominately consist of continuous attributes, while *HeartDisease* has a balanced mix of attributes. Contrary, *BreastCancerOriginal* solely consist of categorical attributes.

Dataset	Attributes			#	Central tendency		Outliers (k=3)		Hopkins
	Cont.	Cate.	Total		Skew.	Kurtos.	Outlier	% of #	
BreastCancer-Original	0	10	10	699	1.39	4.3	120	17%	0.75
ForestFires	9	2	11	517	4.14	66.4	83	16%	0.93
HeartDisease	5	9	14	303	0.80	3.0	46	15%	0.78
PimaIndians	9	1	10	768	1.10	5.7	50	7%	0.86
Winequality-white	11	1	12	4898	1.20	9.5	208	4%	0.96
Frogs	21	3	24	7195	0.59	3.6	77	1%	0.94

Table 5: Low-dimensional datasets

With 0.96 and 0.93 in Hopkins statistic, respectively, *Winequality* and *ForestFires* has a strong cluster tendency whereas *BreastCancerOriginal* and *HeartDisease* has low cluster tendency of 0.75 and 0.78 in Hopkins statistics, respectively. *Winequality* only has 4% of observations classified as outliers, although it has a great average kurtosis of 9.5 and absolute average skewness of 1.2, contrary to *ForestFires* with skewness of 4.14 and 16% as outliers. Thus, for the two datasets with strong cluster tendency, *Winequality* has central tendency while *ForestFires* does not have central tendency, which could indicate multiple clusters. *PimaIndians* has some cluster tendency, with greater central tendency than *BreastCancerOriginal*, *ForestFires* and *HeartDisease*.

With more than 30 dimensions, three datasets are categorized as high-dimensional datasets (Table 6). With 1 and 2 categorical attributes, the datasets predominately consist of continuous attributes. With increasing number of dimensions, an increasing proportion of observations are classified as outliers, confirming the curse of dimensionality. Thus, between 19% and 58% of observations are classified as outliers for datasets, with *GesturePhase* additionally having an extreme kurtosis of 49.3. Although kurtosis is high, skewness is relatively low, somehow indicating a central tendency for *GesturePhase*, characterized by long tails (and many outliers), but mostly non-skew distributions. With Hopkins statistic of 0.98 and 0.96, *BreastCancerDianostic* and *GesturePhase* both have strong cluster tendency.

Dataset	Attributes			#	Central tendency		Outliers (k=3)		Hopkins
	Cont.	Cate.	Total		Skew.	Kurtos.	Outlier	% of #	
BreastCancer-Diagnostic	31	1	32	569	1.85	11.5	125	22%	0.98
GesturePhase	32	1	33	9873	1.64	49.3	5681	58%	0.96
Ionosphere	32	2	34	351	0.61	3.3	65	19%	0.68

Table 6: High-dimensional datasets

4.2. OUTLIER DETECTION AND CLUSTERING ANALYSIS

The syntax for executing the analysis of the selected datasets is as follows:

I - Initial settings

1. Select datasets: *datasets*={*BreastCancerDiagnostic*, *ForestFires*, ... , *Ionosphere*}
2. Select outlier detection methods: *outMethod*={*DB-outlier*, *K-NN*, ... , *SOD*}
3. Select clustering algorithms: *clustAlgorithm*={*Hierarchical*, *K-means*}
4. Select evaluation criteria: *evalCriteria*={ R^2 , *Davies-bouldin*, ... , *Dunn*}

II - Standardize datasets

5. For *i* in *datasets*:
 - a. Standardize dataset: *datasets[i]* <- *std(datasets[i])*
6. Return: *datasets*

III - Apply outlier removal

7. For *i* in *datasets*:
 - a. Remove *n*% random observations: *randomTreatedDataset* <- *remove(dataset, n)*
 - b. For *j* in *outMethod*:
 - i. Apply outlier detection: *treatedDataset[i]* <- *outMethod[j](datasets[i])*
 - ii. Remove top-*n*% outliers
 - c. Return: *treatedDatasets*
 - d. Merge: *treatedDatasets* <- *merge(treatedDatatsets, randomTreatedDataset)*
8. Return: *treatedDatasets*

IV - Apply clustering

9. For *i* in *treatedDatasets*:
 - a. Apply k-means: *kmeansClust[i]* <- *kmeans(treatedDatasets[i])*
 - b. Apply hierarchical: *hierarClust[i]* <- *hierarchical(treatedDatasets[i])*
- Return: *kmeansClust*, *hierarClust*

V - Evaluation

10. For i in $kmeansClust$ & $hierarClust$:
 - a. For j in $evalCriteria$:
 - i. Apply evaluation: $eval[i] <- evalCriteria[j](kmeansClust[i] \& hierarClust[i])$
 - b. Return: $eval$
11. Return: $eval$

4.2.1. Outlier detection settings

Following outlier detection algorithm settings and comments are put in place:

- For DB-outlier the fraction parameter P , labelling an observation outlier or inlier, is omitted and outlierness is judged on the amount of observations within neighborhood, given distance D . D is subjectively set to 20% of the distance between the two most distant observations in the respective dataset DB-outlier is applied to
- For kNN aggregated (abbreviation: KNN_AGG) the aggregation of distance over different kNN 's is set from $k = 10$ to $k = 15$
- For kNN graph (abbreviation: KNN_IN) the k parameter is set to 30 and the threshold T is omitted
- For LOF, the k parameter for the number of neighbors to compare density with is set to 30
- For INFLO, the k parameter for the number of reverse neighbors to compare density with is set to 30
- For COF, the k parameter for the $SBNtrail$ is set to 30
- For LOCI, $MDEF$ is used to judge outlierness. The parameter k is set to 30 and α is set to 0.5 as recommended (Papadimitriou et al., 2003)
- For LOOP, the parameter k is set to 30 and λ is set to 3 as recommended (H-P. Kriegel et al., 2009)
- For ABOD, the parameter k for neighbors to compare angle with is set to 30
- For SOD, the parameter k is set to 30 and α is set to 0.8 as recommended by (H-P. Kriegel et al., 2008)
- For FBOD, the chosen input algorithms are LOF, COF, INFLO and LOOP. As discussed in the literature review, the outputs of the input algorithms are normalized from 0 to 1 and the final output is an average of the normalized outputs. As all the input algorithms are density-based approaches, FBOD is considered a density-based approach in the following presented results.

- For comparison with distance- and density-based approaches, the two statistical approaches Z-score and Mahalanobis distance (abbreviation: MAHA) are selected. The most extreme values in output are considered top-n % outliers.
- With 2 statistical methods, 3 distance-based algorithms, 6 density-based algorithms, 2 high-dimensional algorithms and random removal (RNDM), the total number of inputs are 14.

4.2.2. Clustering settings

Following cluster algorithm settings are configured:

- The span of cluster solutions is set from 1 to 15 clusters
- For k-means clustering, the number of iterations are set to 25 to overcome the vulnerability of the initial seeds. The best result of the 25 iterations is chosen as final output for a given set of clusters
- For hierarchical clustering, the best linkage method is identified by applying all linkage methods to dataset treated with random removal of observations (abbreviation: RNDM). The best linkage outcome proceeds as an input for hierarchical clustering algorithm when applied to datasets treated with outlier removal.

4.3. COHERENCE FOR EVALUATION METRICS

With 14 chosen outlier detection methods for 9 datasets, 3 removal percentages (5-10-15%), 2 clustering algorithms with 1 to 15 clusters, a total of 11.340 cluster solutions are returned. Moreover, for each cluster solution 6 evaluation metrics are returned, making the total number of figures 68.040. To overcome the interpretation issue of the many solutions, the evaluation metrics are averaged out by outlier detection method for each dataset by removal percentage. As Davies-Bouldin, Dunn, Silhouette, cluster correlation and cluster index requires at least two clusters for computation, the outcome is averaged out from 2 to 15 clusters.

To ensure a unified approach for interpretation of cluster results, the returned clustering evaluation metrics are inspected for linear dependence.

For the results to have a coherent causation with respect to clustering objectives, interdependencies are required to be present. That is, if evaluation metric x returns a certain result, evaluation metric y returns a similar result indicating the same result. The required dependencies are:

- R^2 has a negative impact on Davies-Bouldin, cluster correlation and cluster index and a positive impact on Silhouette and Dunn
- Cluster index has a positive impact on Davies-Bouldin and cluster correlation and a negative impact on Silhouette and Dunn
- Cluster correlation has a positive impact on Davies-Bouldin and a negative impact on Silhouette and Dunn.
- Dunn has a negative impact on Davies-Bouldin and a positive impact on Silhouette
- Silhouette has a negative impact on Davies-Bouldin

5. RESULTS AND DISCUSSION

5.1. EVALUATION METRICS

For the chosen datasets undergoing analysis, the evaluation metrics criteria are met to some extent. Correlation matrices for evaluation metrics are available in annex for each dataset.

For *BreastCancerOriginal* dataset, 10 out of 15 relationships for evaluation metrics have a strong correlation⁴ while one relationship has moderate correlation⁵. The remaining 4 relationships have no correlation. All correlations are consistent with necessary dependencies for coherence.

For *ForestFires* dataset, 5 out of 15 relationships for evaluation metrics have a strong correlation, 3 have moderate correlation and one has weak correlation⁶. The remaining 7 relationships have no correlation. Of the 3 relationships with some correlation, the relationship between Dunn and Davies-Bouldin has a correlation of 0.66, violating the negative dependency necessary for coherence.

For *HeartDisease* dataset, 12 out of 15 relationships for evaluation metrics have a strong correlation while the remaining 3 relationships have moderate correlation. All correlations are consistent with necessary dependencies for coherence.

For *PimaIndians* dataset, 2 out of 15 relationships for evaluation metrics have a strong correlation while 8 have moderate correlation and one has weak correlation. The remaining 4 relationships have no correlation. All correlations are consistent with necessary dependencies for coherence.

For *Frogs* dataset, all relationships for evaluation metrics have a strong correlation. All relationships are consistent with necessary dependencies for coherence.

For *BreastCancerDiagnostic* dataset, 5 out of 15 relationships for evaluation metrics have a strong correlation, 5 have moderate correlation and 2 have weak correlation. The remaining 3 relationships have no correlation. Of the 2 relationships with weak correlation, the relationship between R^2 and Silhouette has a correlation of -0.33, violating the positive dependency necessary for coherence. Moreover, R^2 and Davies-Bouldin have a correlation of 0.51, also violating the negative dependency necessary for coherence.

⁴ Strong correlation is defined as $|\text{cor}(x1, x2)| \geq 0.7$

⁵ Moderate correlation is defined as $|\text{cor}(x1, x2)| \geq 0.5$

⁶ Weak correlation is defined as $|\text{cor}(x1, x2)| \geq 0.3$

For *GesturePhase* dataset, 4 out of 15 relationships for evaluation metrics have a strong correlation, 2 have moderate correlation and 2 have weak correlation. The remaining 7 have no correlation. Of the 2 relationships with some correlation, the relationship between cluster index and Dunn has a correlation of 0.54, violating the negative dependency necessary for coherence.

For *Ionosphere* dataset, 14 out of 15 relationships for evaluation metrics have a strong correlation and one has moderate correlation. All relationships are consistent with necessary dependencies for coherence.

For *WineQuality* dataset, 7 out of 15 relationships for evaluation metrics have a strong correlation, one has moderate correlation and one has weak correlation. The remaining 6 have no correlation. All correlations are consistent with necessary dependencies for coherence.

Summarizing the dependencies for evaluation metrics across the returned results, the results are difficult to induce a clear conclusion from. Table 7 highlights the correlation for evaluation metrics when checked for dependence against R^2 .

Dataset	Correlation with R^2					Coherence
	Davies-Bouldin	Dunn	Silhouette	Cluster-Correlation	Cluster index	
HeartDisease	-0.77**	0.71**	0.88**	-0.73**	-0.88**	Coherent
Frogs	-0.91**	0.90**	0.73**	-0.76**	-0.92**	Coherent
Ionosphere	-0.95**	0.74**	0.95**	-0.88**	-0.85**	Coherent
PimaIndians	-0.63**	0.55**	0.60**	-0.39*	-0.53**	Coherent
WineQuality	-0.63**	0.39*	0.72**	-0.71**	-0.78**	Coherent
ForestFires	-0.47*	0.21	0.66**	-0.00	-0.81**	Incoherent
BreastCancerOriginal	-0.77**	0.45*	0.49**	-0.12	-0.73**	Incoherent
BreastCancerDiagnostics	0.51**	0.67**	-0.34*	0.03	0.09	Incoherent
GesturePhase	-0.88**	0.50**	-0.08	0.19	0.24	Incoherent

Table 7: Evaluation metrics correlation with R^2 . **0.01 significance, *0.05 significance

For the datasets *HeartDisease*, *Frogs*, *Ionosphere*, *PimaIndians* and *WineQuality* a coherence for all evaluation metrics is achieved with strong or some correlations present. In particular the evaluation metrics for *Frogs* and *Ionosphere* have strong dependencies with R^2 , meaning when R^2 indicates better clustering results, other evaluation metrics also indicates better clustering results with respect to defined clustering objectives.

For the datasets *ForestFires*, *BreastCancerOriginal*, *BreastCancerDiagnostic* and *GesturePhase* a poor almost non-existing coherence is observed. One evaluation metric can indicate better clustering results with respect to clustering objectives, while other evaluation metrics indicates worse clustering results. Thus, it cannot be deduced that there is a relationship between the different evaluation metrics for these dataset, making it impossible to have a unified interpretation approach for these datasets.

For the following analysis of the results, incoherent cluster results are excluded as it is cannot be determined if outlier detection has improved clustering results when evaluation metrics are conflicting. Comments and discussion for the incoherent cluster results are found in annex. For the coherent cluster results, the results are inspected from the perspective of R^2 as an indication of improved results for all evaluation metrics.

5.2. HEARTDISEASE DATASET

Considering the hierarchical clustering results for the *HeartDisease* dataset (Table 8), the greatest improvement of clustering results is achieved with COF treated dataset (15%), returning a R^2 of 0.414, 0.041 points better than RNDM treated dataset.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	SOD	KNN_IN	COF	0.389	0.400	0.414	2378	2076	1926
2	LOF	COF	FBOD	0.388	0.399	0.413	2320	2171	1869
3	COF	FBOD	LOF	0.388	0.398	0.411	2386	2096	1888
4	LOOP	LOCI	KNN_IN	0.388	0.397	0.407	2324	2177	1888
5	FBOD	LOF	LOOP	0.388	0.397	0.406	2324	2099	1917
6	KNN_IN	LOOP	LOCI	0.385	0.396	0.402	2342	2113	2031
7	LOCI	SOD	SOD	0.384	0.391	0.402	2402	2219	2015
8	INFLO	INFLO	INFLO	0.382	0.389	0.400	2361	2142	1945
9	KNN_AGG	DB	KNN_AGG	0.380	0.384	0.397	2335	2096	1873
10	DB	KNN_AGG	MAHA	0.376	0.380	0.385	2341	2113	2094
11	ZSCORE	RNDM	DB	0.370	0.378	0.382	2387	2383	1891
12	RNDM	ZSCORE	ZSCORE	0.368	0.370	0.375	2556	2168	1962
13	MAHA	MAHA	RNDM	0.367	0.369	0.373	2481	2311	2295
14	ABOD	ABOD	ABOD	0.360	0.363	0.363	2448	2273	2117

Table 8: Average results for hierarchical clustering applied to the *HeartDisease* dataset

The seven best solutions are predominately returned by the density-based approaches, LOF, LOCI, FBOD, LOOP and COF, the distance-based approach KNN_IN and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.010 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD, the distance-based approaches DB and KNN_AGG and the density-based methods INFLO. Of these solutions, MAHA and ABOD returns worse results when compared to RNDM treated dataset. Furthermore, R^2 decrease with 0.005 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 1873 (15%) followed by DB and LOOP with an SSW of 1891 and 1917, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.397, 0.382 and 0.406, respectively.

For k-means clustering (Table 9), the greatest improvement of clustering results is achieved with COF treated dataset (15%), returning an R^2 of 0.432, 0.037 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, COF, LOF, LOCI, FBOD and LOOP, the distance-based approach KNN_IN and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.009 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD, the distance-based approaches DB and KNN_AGG and the density-based methods INFLO. Of these solutions, ZSCORE and ABOD returns worse results when compared to RNDM treated dataset. Furthermore, R^2 increase with 0.003 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 1808 (15%) followed by KNN_AGG and KNN_IN with an SSW of 1820 and 1822, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.409, 0.414 and 0.428, respectively.

For *HeartDisease* it is concluded that k-means clustering returns the best results, with an R^2 of 0.432 compared to an R^2 of 0.414 for hierarchical clustering. In general, the clustering results are poor, returning the second lowest R^2 for all clustering results across datasets.

When asked how outlier detection can improve clustering results, the greatest impact is observed for hierarchical clustering, improving the overall clustering results with 0.039 points, compared to 0.037 for k-means clustering (Figure 16). Furthermore, R^2 increase by 0.01 for hierarchical clustering, compared to 0.009 for k-means, when removing additional 5% observations.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	COF	COF	COF	0.411	0.422	0.432	2298	2087	1867
2	FBOD	LOOP	LOF	0.408	0.421	0.431	2252	2027	1823
3	SOD	SOD	KNN_IN	0.407	0.415	0.428	2306	2130	1822
4	LOOP	KNN_IN	LOOP	0.407	0.413	0.425	2252	2032	1857
5	KNN_IN	LOF	LOCI	0.407	0.412	0.423	2259	2047	1962
6	KNN_AGG	FBOD	FBOD	0.406	0.412	0.422	2239	2050	1844
7	LOF	LOCI	SOD	0.405	0.409	0.420	2257	2134	1956
8	INFLO	KNN_AGG	INFLO	0.401	0.407	0.417	2289	2023	1892
9	DB	INFLO	KNN_AGG	0.400	0.406	0.414	2251	2084	1820
10	LOCI	DB	DB	0.400	0.404	0.409	2340	2026	1808
11	MAHA	MAHA	MAHA	0.396	0.398	0.405	2368	2205	2026
12	RNDM	RNDM	RNDM	0.393	0.396	0.395	2432	2333	2176
13	ZSCORE	ZSCORE	ZSCORE	0.392	0.392	0.392	2305	2092	1909
14	ABOD	ABOD	ABOD	0.385	0.383	0.384	2351	2198	2046

Table 9: Average results for k-means clustering applied to the HeartDisease dataset

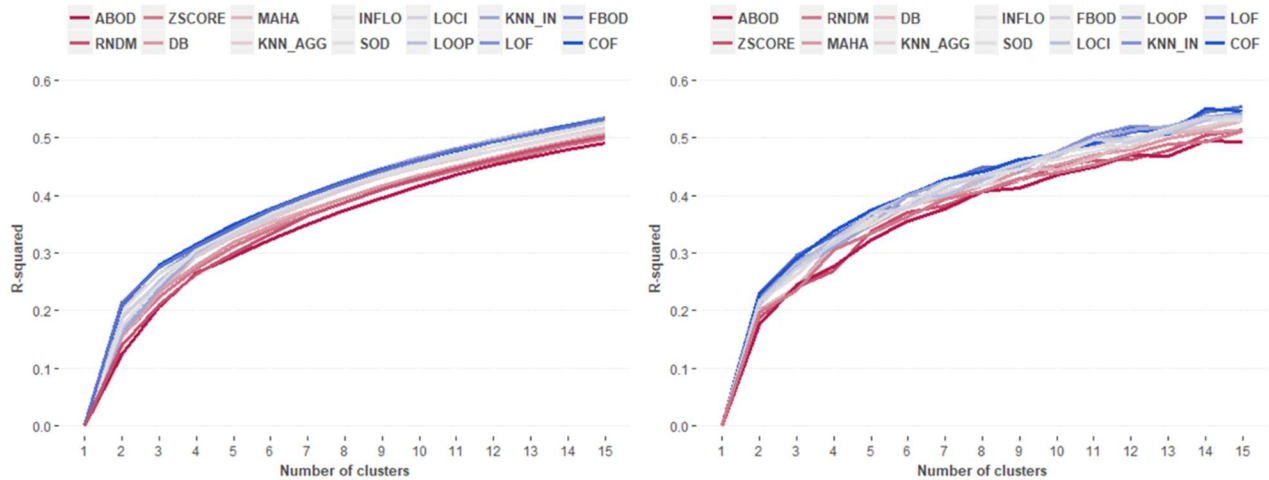


Figure 16: R^2 for HeartDisease with hierarchical clustering (left) and k-means (right) with 15% removed observations

Both ABOD and ZSCORE, being the two worst outlier detection method for *HeartDisease*, have just a slightly worse R^2 when compared to RNDM, meaning a better outcome is achieved by removing random observations.

Inspecting SSW, the picture differs between the two clustering algorithms. For hierarchical clustering the lowest SSW is observed for LOF and FBOD, while for k-means clustering the lowest SSW is observed for DB and KNN_AGG. Thus, KNN_AGG and DB creates the densest clusters in absolute figures but differ in ability to explain the proportion of remaining variance.

Inspecting evaluation metrics in annex, COF (15%) has the strongest cluster correlation and the greatest Silhouette for hierarchical clustering, while having the second lowest cluster index and the third greatest Silhouette for k-means clustering.

5.3. PIMA INDIANS DATASET

For hierarchical clustering, the greatest improvement of clustering results in the *PimaIndians* dataset is achieved with COF treated dataset (15%), returning an R^2 of 0.486, 0.045 points better than RNDM treated dataset (Table 10).

The seven best solutions are predominately achieved with the density-based approaches, LOF, COF, FBOD, LOCI and LOOP and the distance-based approaches KNN_AGG and KNN_IN. For these solutions, R^2 increase with 0.008 in average when removing additional 5% observations. That is, relatively less variance is explained.

The seven worst solutions are returned by the distance-based approach DB, the high-dimensional approaches ABOD and SOD, the statistical approaches ZSCORE and MAHA and the density-based approaches INFLO and LOCI. Of these solutions, MAHA and ABOD return worse results than with RNDM treated dataset. Furthermore, for these solutions R^2 increase with 0.002 points in average when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 2032 (15%) followed by KNN_AGG and ZSCORE with an SSW of 2051 and 2067, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R^2 of 0.445, 0.464 and 0.446, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	KNN_IN	COF	COF	0.463	0.479	0.486	2808	2468	2228
2	COF	FBOD	KNN_IN	0.462	0.466	0.481	2807	2349	2128
3	KNN_AGG	LOOP	FBOD	0.460	0.465	0.475	2753	2404	2088
4	LOF	KNN_IN	LOOP	0.460	0.462	0.472	2723	2441	2146
5	LOOP	KNN_AGG	LOF	0.458	0.462	0.471	2764	2350	2089
6	LOCI	LOCI	LOCI	0.452	0.459	0.471	2796	2431	2148
7	FBOD	LOF	KNN_AGG	0.451	0.458	0.464	2789	2395	2051
8	MAHA	SOD	SOD	0.448	0.450	0.463	2911	2796	2490
9	ZSCORE	INFLO	INFLO	0.448	0.449	0.458	2763	2558	2287
10	DB	MAHA	ZSCORE	0.446	0.445	0.446	2750	2691	2067
11	SOD	ZSCORE	DB	0.444	0.444	0.445	3063	2377	2032
12	INFLO	DB	RNDM	0.442	0.438	0.441	2946	2386	2917
13	ABOD	RNDM	MAHA	0.435	0.436	0.441	2890	3078	2499
14	RNDM	ABOD	ABOD	0.434	0.425	0.439	3313	2657	2352

Table 10: Average results for hierarchical clustering applied to the PimaIndians dataset

For k-means clustering, the greatest improvement of clustering results is achieved with COF treated dataset (15%), returning an R² of 0.524, 0.046 points better than RNDM treated dataset (Table 11).

The seven best solutions are predominately achieved with the density-based approaches, LOF, COF, FBOD, LOCI and LOOP and the distance-based approach KNN_AGG and KNN_IN. For these solutions, R² increase with 0.008 in average, when increasing the percentage of removed observations with 5%. That is, relatively less variance is explained.

The seven worst solutions are returned by the distance-based approach DB, the high-dimensional approach ABOD and SOD, the statistical approaches ZSCORE and MAHA and the density-based approach INFLO. Of these solutions, DB return worse results than with RNDM treated dataset. Furthermore, for these solutions R² increase with 0.002 points in average, when increasing the percentage of removed observations with 5%. That is, relatively less variance is explained.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 1884 (15%) followed by LOCI and KNN_AGG with an SSW of 1986 and 1897, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R² of 0.485, 0.511 and 0.504, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	KNN_IN	COF	COF	0.499	0.511	0.524	2622	2315	2064
2	COF	LOOP	LOOP	0.497	0.505	0.516	2624	2222	1965
3	LOF	FBOD	KNN_IN	0.494	0.504	0.512	2549	2250	2003
4	KNN_AGG	KNN_IN	LOCI	0.493	0.503	0.511	2586	2259	1986
5	LOOP	KNN_AGG	FBOD	0.493	0.502	0.507	2587	2176	2000
6	LOCI	LOF	LOF	0.492	0.502	0.507	2588	2203	1948
7	INFLO	LOCI	KNN_AGG	0.488	0.501	0.504	2704	2245	1897
8	FBOD	INFLO	INFLO	0.487	0.490	0.493	2609	2365	2135
9	SOD	SOD	SOD	0.485	0.487	0.491	2835	2610	2362
10	ZSCORE	MAHA	ZSCORE	0.483	0.485	0.487	2585	2500	1914
11	MAHA	ZSCORE	DB	0.480	0.484	0.485	2742	2206	1884
12	DB	ABOD	MAHA	0.478	0.482	0.483	2592	2252	2310
13	ABOD	RNDM	ABOD	0.476	0.480	0.479	2650	2895	2008
14	RNDM	DB	RNDM	0.471	0.479	0.478	3115	2211	2691

Table 11: Average results for k-means clustering applied to the PimaIndians dataset

For *PimaIndians*, it is concluded that k-means clustering returns the best results, with an R² of 0.524 compared to an R² of 0.486 for hierarchical clustering, a difference of 0.038 points.

When asked how outlier detection can improve clustering results, the greatest impact is observed for k-means clustering, improving the overall clustering results with 0.046 points, compared to 0.045 for hierarchical clustering (Figure 17). However, R² only increase with 0.008 in average for both hierarchical clustering and k-means, when removing additional 5% observations.

Both ABOD and MAHA, being the two worst outlier detection method with hierarchical clustering, have a slightly worse R² when compared to RNDM, meaning a better outcome is achieved by removing random observations.

Inspecting SSW, it is concluded that KNN_AGG, DB and ZSCORE all excel in creating dense clusters, with the lowest SSW but fail to remove outliers between clusters, leading to a mediocre R² when compared to density-based approaches such as LOOP, LOF and COF and the distance-based approach KNN_IN. Thus, KNN_AG, DB and ZSCORE are best at reducing the overall sum of squares but fails to explain the remaining variance.

Inspecting evaluation metrics in annex, COF (15%) has the greatest Silhouette, the lowest Davies-Bouldin and the lowest cluster index for hierarchical clustering, while having the lowest cluster index, the second greatest Silhouette and second lowest Davies-Bouldin for k-means clustering.

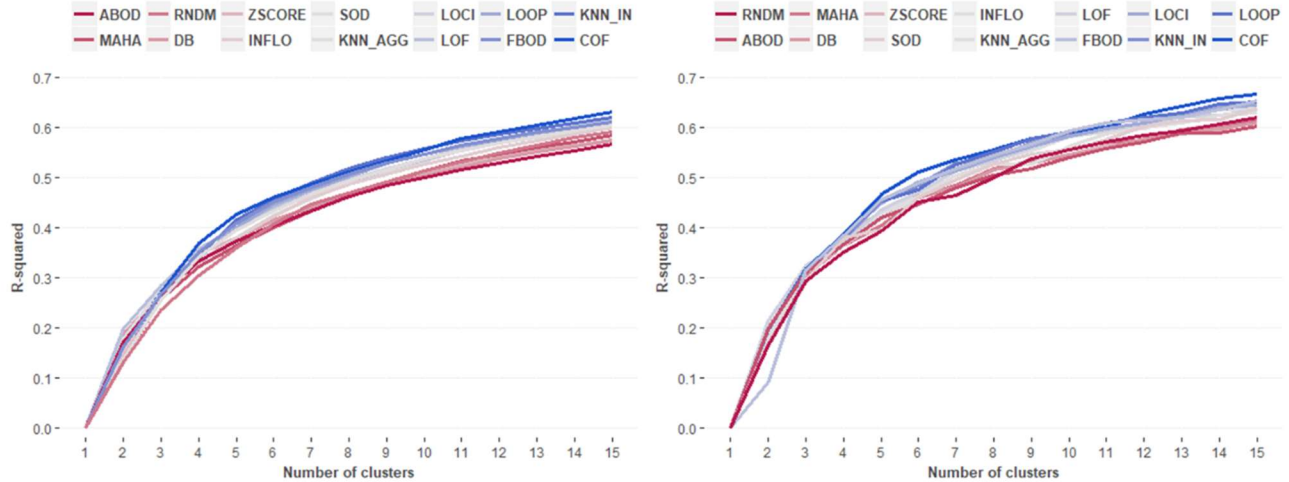


Figure 17: R^2 for PimaIndians with hierarchical clustering (left) and k-means (right) with 15% removed observations

5.4. FROGS DATASET

The greatest improvement of hierarchical clustering results in the *Frogs* dataset is achieved with FBOD treated dataset (15%), returning an R^2 of 0.694, 0.093 points better than RNDM treated dataset (Table 12).

The seven best solutions are predominately achieved with the density-based approaches, LOF, COF, FBOD, LOCI and LOOP and the distance-based approaches KNN_AGG and DB. For these solutions, R^2 increase with 0.018 in average, when increasing the percentage of removed observations with 5%. That is, relatively less variance is explained.

The seven worst solutions are returned by the distance-based approach KNN_IN, the high-dimensional approaches ABOD and SOD, the statistical approaches ZSCORE and MAHA and the density-based approach INFLO. Of these solutions, MAHA returns worse results than with RNDM treated dataset. Furthermore, for these solutions R^2 increase with 0.010 points in average, when increasing the percentage of removed observations with 5%. That is, relatively less variance is explained.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 10118 (15%) followed by ZSCORE and KNN_AGG with an SSW of 10396 and 10428, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R^2 of 0.679, 0.673 and 0.693, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	DB	DB	FBOD	0.656	0.685	0.694	14122	11456	11045
2	KNN_AGG	KNN_AGG	KNN_AGG	0.654	0.684	0.693	14427	11933	10428
3	FBOD	LOCI	LOF	0.648	0.678	0.693	15036	12690	11121
4	LOOP	COF	COF	0.648	0.677	0.685	14901	12729	11561
5	LOCI	FBOD	LOOP	0.648	0.677	0.683	15122	12771	11526
6	COF	LOF	LOCI	0.647	0.676	0.680	15014	12765	11585
7	SOD	LOOP	DB	0.643	0.676	0.679	14953	12761	10118
8	LOF	SOD	KNN_IN	0.643	0.669	0.676	15315	12677	11956
9	ZSCORE	ZSCORE	ZSCORE	0.641	0.668	0.673	14791	12069	10396
10	KNN_IN	KNN_IN	SOD	0.636	0.664	0.671	15516	13322	11568
11	ABOD	ABOD	INFLO	0.623	0.639	0.644	16086	14165	13265
12	INFLO	INFLO	ABOD	0.621	0.634	0.640	16358	14731	12848
13	RNDM	MAHA	MAHA	0.602	0.610	0.609	18221	16753	15782
14	MAHA	RNDM	RNDM	0.599	0.608	0.601	18320	16873	16227

Table 12: Average results for hierarchical clustering applied to the Frogs dataset

For k-means clustering, the greatest improvement of clustering results is achieved with DB treated dataset (15%), returning an R² of 0.663, 0.074 points better than RNDM treated dataset (Table 13).

The seven best solutions are predominately achieved with the density-based approaches, LOF, COF, FBOD and LOCI, the distance-based approaches KNN_AGG and DB, the statistical test ZSCORE and the high-dimensional approaches ABOD and SOD. For these solutions, R² increase with 0.010 in average, when removing additional 5% observations. That is, relatively less variance is explained.

The seven worst solutions are predominately returned by the distance-based approach KNN_IN, the statistical approach MAHA and the density-based approach INFLO. Of these solutions, MAHA returns worse results than with RNDM treated dataset. Furthermore, for these solutions R² increase with 0.006 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 10606 (15%) followed by ZSCORE and ABOD with an SSW of 11434 and 11779, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R² of 0.663, 0.640 and 0.635, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	ZSCORE	LOF	DB	0.628	0.654	0.663	15362	13634	10606
2	FBOD	KNN_AGG	LOF	0.627	0.645	0.656	15859	13388	12442
3	DB	FBOD	FBOD	0.626	0.639	0.650	15355	14304	12858
4	LOCI	DB	ZSCORE	0.624	0.639	0.640	16131	13130	11434
5	LOF	KNN_IN	KNN_AGG	0.622	0.638	0.636	16210	14375	12354
6	ABOD	SOD	ABOD	0.621	0.634	0.635	15501	13987	11779
7	KNN_IN	LOOP	LOCI	0.618	0.632	0.634	16298	14476	13255
8	SOD	COF	COF	0.615	0.629	0.630	16146	14630	13601
9	LOOP	ZSCORE	KNN_IN	0.613	0.627	0.628	16392	13572	13725
10	INFLO	LOCI	SOD	0.610	0.627	0.626	16823	14690	13134
11	COF	ABOD	LOOP	0.607	0.617	0.620	16747	13942	13829
12	KNN_AGG	INFLO	INFLO	0.596	0.613	0.617	16856	15606	14241
13	RNDM	RNDM	MAHA	0.590	0.591	0.591	18597	17799	16517
14	MAHA	MAHA	RNDM	0.584	0.591	0.589	19035	17599	16594

Table 13: Average results for k-means clustering applied to the Frogs dataset

For *Frogs* it is concluded that hierarchical clustering returns the best results, with an R² of 0.694 compared to an R² of 0.663 for k-means clustering, making it the second-best clustering results for all datasets.

When asked how outlier detection can improve clustering results (Figure 18), the greatest impact is also observed for hierarchical clustering, improving the overall clustering results with 0.093 points compared to 0.074 for hierarchical clustering. Furthermore, the greatest increase in explained variance is also for hierarchical clustering, with an average increase of 0.018 points, when removing additional 5% observations.

Interestingly, the best results for hierarchical clustering are predominantly returned by the two distance-based approaches DB and KNN_AGG, both in terms of explained variance and absolute figures. For k-means clustering, ZSCORE produces a great solution when only 5% of observations are removed, indicating that extreme outliers exist in the tails of the distributions easily detectable for ZSCORE. With 10% and 15% removed observations, the results for ZSCORE (k-means) decrease relatively to other methods and the two density-based approaches LOF and FBOD are among the best solutions. This indicates that more outliers are removed between the clusters.

Inspecting evaluation metrics in annex, FBOD (15%) has the greatest Silhouette, the lowest Davies-Bouldin and the lowest cluster index for hierarchical clustering, while DB has the greatest Dunn, the third lowest cluster index and the third strongest cluster correlation for k-means clustering.

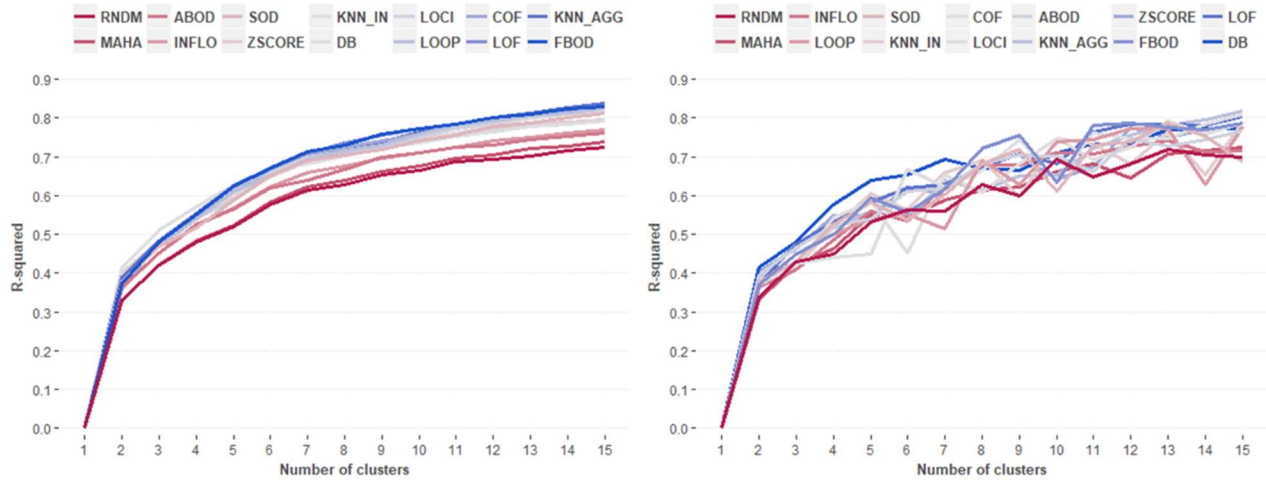


Figure 18: R^2 for Frogs with hierarchical clustering (left) and k-means (right) with 15% removed observations

5.5. IONOSPHERE DATASET

For hierarchical clustering, the greatest improvement of clustering results in the *Ionosphere* dataset is achieved with the KNN_AGG treated dataset (15%), returning an R^2 of 0.604, 0.17 points better than RNDM treated dataset (Table 14).

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD, LOOP, INFLO and LOCI and the distance-based approaches KNN_IN and KNN_AGG and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.053 points in average, when removing additional 5% observations. That is, relatively more variance is explained.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD and the distance-based method DB. Of these solutions, none returns worse results when compared to RNDM treated dataset. Furthermore, for these solutions R^2 increase with 0.03 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 2733 (15%) followed by KNN_IN and SOD with an SSW of 2832 and 3008, respectively. The solutions almost perform equal for explaining variance, with an R^2 of 0.604, 0.594 and 0.583, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	KNN_AGG	KNN_AGG	KNN_AGG	0.483	0.538	0.604	4986	3748	2733
2	LOCI	KNN_IN	KNN_IN	0.478	0.532	0.594	5211	3915	2832
3	INFLO	LOCI	LOCI	0.477	0.532	0.584	5212	4036	3081
4	KNN_IN	INFLO	SOD	0.475	0.530	0.583	5191	3932	3008
5	COF	SOD	DB	0.475	0.525	0.580	5231	3999	3068
6	LOF	COF	LOF	0.472	0.520	0.572	5360	4211	3281
7	SOD	FBOD	FBOD	0.471	0.516	0.565	5155	4315	3326
8	DB	LOF	LOOP	0.470	0.516	0.560	5326	4331	3435
9	FBOD	DB	COF	0.469	0.515	0.560	5424	4238	3437
10	LOOP	ZSCORE	ZSCORE	0.468	0.510	0.558	5451	4182	3254
11	ZSCORE	LOOP	ABOD	0.468	0.509	0.529	5248	4468	3601
12	ABOD	ABOD	INFLO	0.463	0.498	0.519	5311	4292	3511
13	MAHA	MAHA	MAHA	0.454	0.470	0.489	5633	5058	4395
14	RNDM	RNDM	RNDM	0.427	0.436	0.434	6396	5886	5550

Table 14: Average results for hierarchical clustering applied to the Ionosphere dataset

For k-means clustering the greatest improvement of clustering results is achieved with the KNN_AGG treated dataset (15%), returning an R² of 0.604, 0.163 points better than RNDM treated dataset (Table 15).

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD, INFLO and LOCI and the distance-based approaches KNN_IN and KNN_AGG and the high-dimensional approach SOD. For these solutions, R² increase with 0.0514 in average, when removing additional 5% observations. That is, relatively more variance is explained.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD and the distance-based method DB. Of these solutions, none returns worse results when compared to RNDM treated dataset. Furthermore, for these solutions R² increase with 0.028 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 2753 (15%) followed by KNN_IN and SOD with an SSW of 2852 and 2993, respectively. The solutions almost perform equal for explaining variance, with an R² of 0.601, 0.591 and 0.585, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOCI	KNN_AGG	KNN_AGG	0.484	0.543	0.601	5153	3706	2753
2	KNN_AGG	KNN_IN	KNN_IN	0.483	0.537	0.591	4987	3874	2852
3	COF	LOCI	SOD	0.481	0.532	0.585	5168	4040	2993
4	INFLO	INFLO	LOCI	0.480	0.529	0.583	5186	3934	3090
5	LOF	SOD	DB	0.478	0.526	0.577	5301	3996	3094
6	FBOD	FBOD	LOF	0.477	0.522	0.573	5316	4295	3268
7	SOD	LOF	ZSCORE	0.476	0.521	0.565	5103	4285	3199
8	KNN_IN	COF	FBOD	0.475	0.520	0.561	5197	4213	3402
9	ABOD	DB	LOOP	0.473	0.519	0.559	5212	4195	3444
10	DB	ZSCORE	COF	0.472	0.512	0.551	5304	4162	3506
11	LOOP	LOOP	ABOD	0.472	0.506	0.530	5407	4495	3592
12	ZSCORE	ABOD	INFLO	0.469	0.502	0.526	5240	4252	3460
13	MAHA	MAHA	MAHA	0.461	0.469	0.495	5559	5059	4345
14	RNDM	RNDM	RNDM	0.434	0.431	0.438	6209	5938	5575

Table 15: Average results for k-means clustering applied to the *Ionosphere* dataset

For *Ionosphere* it is concluded that hierarchical clustering returns the best results, with an R^2 of 0.604 compared to an R^2 of 0.601 for k-means clustering. The margins between the two clustering algorithms are however very small.

When asked how outlier detection can improve clustering results (Figure 19), the greatest impact is also observed for hierarchical clustering, improving the overall clustering results with 0.17 compared to 0.163 for k-means. As the only dataset, the average proportion of explained variance when removing additional 5% observations, is greater than 5%, meaning a relatively better clustering result is achieved by removing more additional observations. The great improvements for clustering results, regardless of applied clustering method is indicating a lot of outliers, easily identifiable by most outlier detection methods.

Inspecting SSW, it is concluded that KNN_AGG and KNN_IN both excel in creating dense clusters, with the lowest SSW observed for both clustering methods, while also having the greatest R^2 . Thus, KNN_AGG and KNN_IN creates the densest clusters in absolute figures, but also manage to explain the greatest proportion of the remaining variance. That is, being great at minimizing intra-cluster distance and maximizing inter-cluster distance. Returning to statement of meaninglessness of neighborhood in high-dimensional dataset (H-P. Kriegel, Schubert, & Zimek, 2008), the outcome for *Ionosphere* indicates that the neighborhood is still an important property, although working with 34 dimensions.

Inspecting evaluation metrics in appendix, KNN_AGG (15%) has the greatest Silhouette, the second lowest Davies-Bouldin, the greatest Dunn and the strongest cluster correlation for hierarchical clustering, while having the lowest Davies-Bouldin, the greatest Dunn and the strongest cluster correlation for k-means clustering.

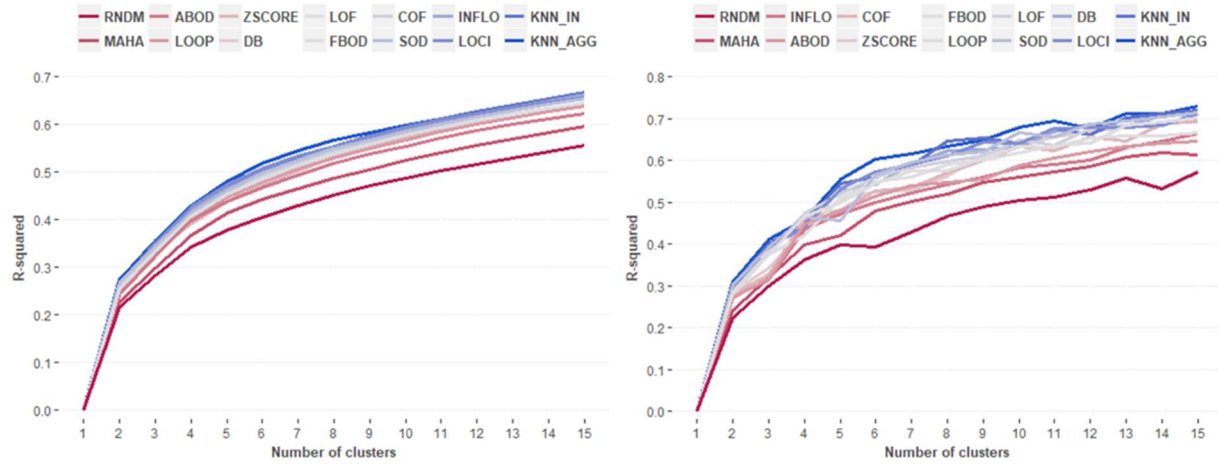


Figure 19: R^2 for Ionosphere with hierarchical clustering (left) and k-means clustering (right) with 15% removed observations

5.6. WINEQUALITY DATASET

For hierarchical clustering, the greatest improvement of clustering results in the *WineQuality* dataset is achieved with the KNN_AGG treated dataset (15%), returning an R^2 of 0.418, 0.045 points better than RNDM treated dataset (Table 16).

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD, LOOP and LOCI and the distance-based approaches KNN_IN and KNN_AGG. For these solutions, R^2 increase with 0.01 points in average, when increasing the percentage of removed observations with 5%. That is, relatively less variance is explained.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD and the distance-based method DB. Of these solutions, none returns worse results when compared to RNDM treated dataset. Furthermore, for these solutions R^2 decrease with 0.005 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 9193 (15%) followed by DB and ZSCORE with an SSW of 9544 and 9683, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R^2 of 0.418, 0.372 and 0.382, respectively.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOF	KNN_IN	KNN_AGG	0.397	0.408	0.418	12353	11265	9193
2	KNN_IN	FBOD	FBOD	0.395	0.406	0.414	12656	10845	9710
3	COF	COF	KNN_IN	0.393	0.406	0.412	12795	11519	10182
4	FBOD	LOOP	LOOP	0.391	0.403	0.411	12416	11029	9960
5	LOCI	LOCI	LOF	0.390	0.402	0.410	12513	11152	9732
6	LOOP	LOF	LOCI	0.388	0.401	0.407	12600	10929	10197
7	SOD	KNN_AGG	COF	0.383	0.389	0.401	13249	10669	10669
8	ZSCORE	INFLO	SOD	0.380	0.387	0.391	12131	11893	11139
9	KNN_AGG	SOD	ZSCORE	0.380	0.385	0.382	12200	12141	9683
10	DB	ZSCORE	INFLO	0.379	0.381	0.381	12052	10813	10951
11	INFLO	DB	RNDM	0.378	0.377	0.373	13252	10637	12962
12	ABOD	MAHA	DB	0.372	0.373	0.372	12628	13248	9544
13	RNDM	RNDM	ABOD	0.372	0.371	0.363	14311	13637	10475
14	MAHA	ABOD	MAHA	0.368	0.371	0.360	14121	11330	12664

Table 16: Average results for hierarchical clustering applied to the WineQuality dataset

For k-means clustering (Table 17), the greatest improvement of clustering results is achieved with the KNN_IN treated dataset (15%), returning an R² of 0.456, 0.04 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD, LOOP and LOCI and the distance-based approaches KNN_IN and KNN_AGG. For these solutions, R² increase with 0.006 points in average, when removing additional 5% observations. That is, relatively less variance is explained.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approaches ABOD and SOD and the distance-based method DB. Of these solutions, none returns worse results when compared to RNDM treated dataset. Furthermore, for these solutions R² increase with 0.0008 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 8616 (15%) followed by DB and ZSCORE with an SSW of 8727 and 8875, respectively. However, the solutions differ in ability to explain the remaining variance, of with an R² of 0.456, 0.25 and 0.433, respectively.

For *WineQuality* it is concluded that k-means clustering returns the best results, with an R² of 0.456 compared to an R² of 0.418 for hierarchical clustering, a difference of 0.038 points.

Rank	Method			R ²			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	COF	COF	KNN_IN	0.442	0.447	0.456	11778	10722	9414
2	FBOD	LOF	KNN_AGG	0.439	0.446	0.454	11557	10104	8613
3	LOF	FBOD	LOF	0.438	0.445	0.453	11509	10317	9019
4	KNN_IN	LOOP	COF	0.438	0.445	0.452	11768	10259	9759
5	LOOP	KNN_IN	FBOD	0.434	0.444	0.452	11653	10568	9268
6	ZSCORE	KNN_AGG	LOOP	0.432	0.443	0.450	11123	9724	9303
7	DB	LOCI	LOCI	0.430	0.439	0.433	11058	10474	9737
8	LOCI	ZSCORE	ZSCORE	0.430	0.435	0.433	11689	9864	8875
9	KNN_AGG	INFLO	INFLO	0.428	0.430	0.432	11257	11053	10059
10	INFLO	DB	SOD	0.424	0.430	0.428	12258	9739	10463
11	SOD	SOD	DB	0.423	0.423	0.425	12373	11390	8727
12	RNDM	RNDM	ABOD	0.420	0.422	0.424	13249	12340	9143
13	MAHA	ABOD	RNDM	0.419	0.419	0.416	12978	10341	11745
14	ABOD	MAHA	MAHA	0.416	0.417	0.414	11611	12317	11602

Table 17: Average results for k-means clustering applied to the WineQuality dataset

When asked how outlier detection can improve clustering results, the greatest impact is observed for hierarchical clustering, improving the overall clustering results with 0.045 compared to 0.04 for k-means (Figure 20). Furthermore, the greatest increase in explained variance is also for hierarchical clustering, with an average increase of 0.01 points, when removing additional 5% observations.

Both ABOD and MAHA, being the two worst outlier detection method with hierarchical clustering, have a slightly worse R² when compared to RNDM, meaning a better outcome is achieved by removing random observations.

Inspecting SSW, it is concluded that KNN_AGG, DB and ZSCORE all excel in creating dense clusters, with the lowest SSW observed for both clustering methods. Thus, with the greatest R² and the lowest SSW, KNN_AGG creates the densest clusters in absolute figures, but also manage to explain the greatest proportion of the remaining variance for hierarchical clustering.

Inspecting evaluation metrics in appendix, KNN_AGG (15%) has the strongest cluster correlation, the greatest Dunn index and the third greatest Silhouette for hierarchical clustering. KNN_IN has the lowest cluster index and the greatest Silhouette for k-means clustering. That is, being great at minimizing intra-cluster distance and maximizing inter-cluster distance.

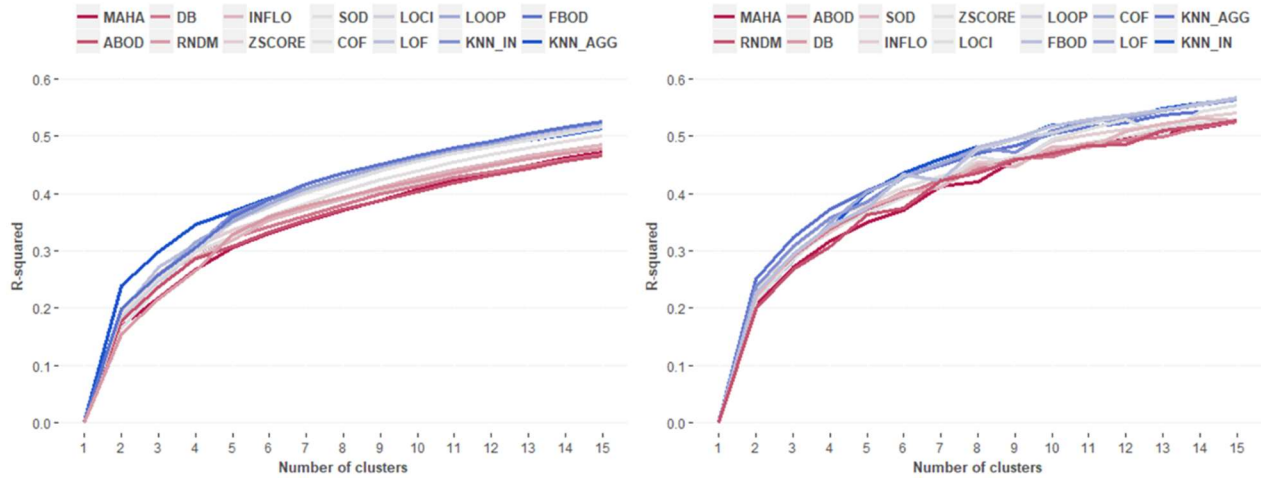


Figure 20: R^2 for WineQuality with hierarchical clustering (left) and k-means clustering (right) with 15% removed observations

5.7. OVERALL PERFORMANCE OF OUTLIER ALGORITHMS

5.7.1. Hierarchical clustering

For coherent results with hierarchical clustering, a mix of the two distance-based approaches KNN_AGG and KNN_IN and the density-based approaches LOOP, FBOD, LOCI, COF and LOF are observed to have the greatest impact on clustering results (Figure 21). For two out of five datasets, hierarchical clustering has the best overall clustering outcome for R^2 , although having the greatest improvement for four out of five datasets.

With 18.5% of ranking 1, LOCI and KNN_AGG have the greatest proportion of best performances, followed by SOD with 14.8% and COF and KNN_IN with 11,1% both. However, inspecting the margins between ranking 1 and 7, a small margin of 0.01 to 0.02 points in average is observed for datasets *WineQuality* and *HeartDisease*.

The statistical approaches ZSCORE and MAHA have a great variance in results, observed as the best and worst solutions. Especially MAHA has poor performance, making up 33% of ranking 13 and is often observed as a worse solution than removing random observations.

The high-dimensional approach ABOD accounts for 44% of ranking 14, making it a worse solution compared to RNDM with a proportion of 25% of ranking 14.

Moreover, the density-based approach INFLO generally performs poorly and has strong deviation from other density-based approaches. The distance-based approach DB also has great variance in results, being both the best and worst performing method.

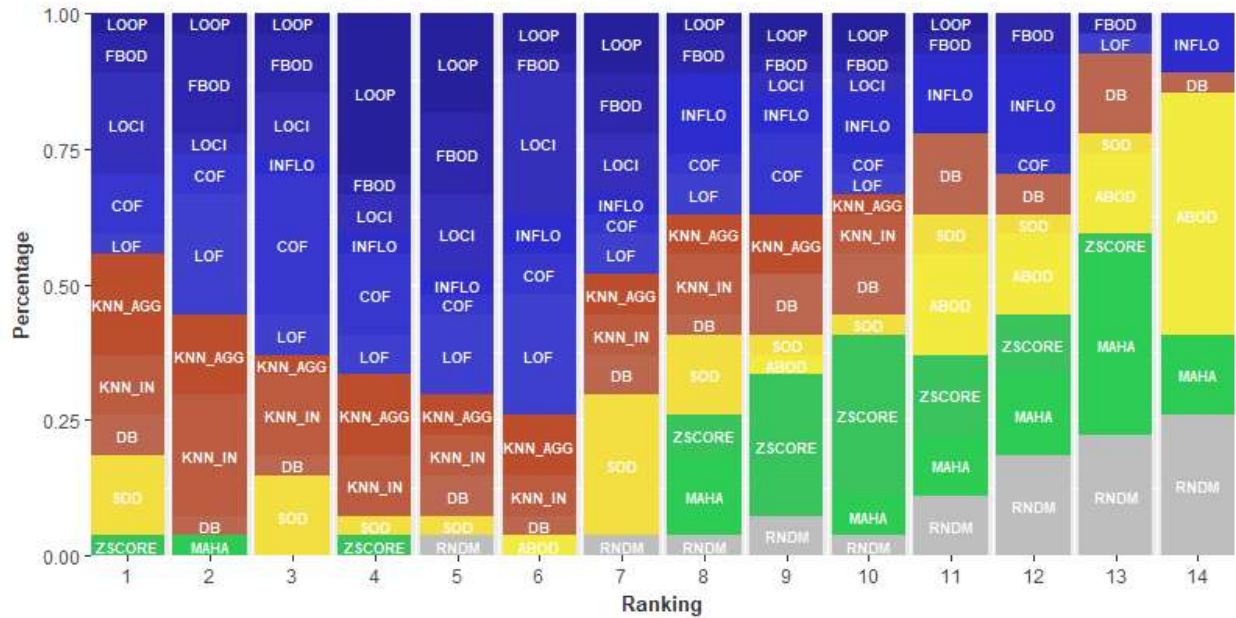


Figure 21: Proportion of outlier methods ranking for hierarchical clustering across all coherent solutions

5.7.2. K-means clustering

For coherent results with k-means clustering (Figure 22), the proportion of density-based approaches is increased when compared to ranking 1 and ranking 2 of hierarchical clustering. For three out of five datasets, k-means has the best overall clustering outcome for R^2 , making it the best clustering tool although having less improvement than hierarchical clustering.

With 33% of ranking 1, k-means returns significantly better results for COF when compared to same dataset for hierarchical clustering. The two distance-based approaches KNN_IN and KNN_AGG accounts for 22% of ranking 1, compared to 29% of ranking 1 for hierarchical clustering.

Interestingly, the distribution of rankings for RNDM has less variance when compared to hierarchical clustering. For ranking 12, 13 and 14 RNDM accounts for 29%, 18% and 44%, respectively, compared to 18%, 22% and 25% for hierarchical clustering. This suggest that k-means has better properties to utilize the outlier treated dataset.

Again, the statistical approaches ZSCORE and MAHA have a great variance in results, observed as the best and worst solutions and the high-dimensional approach ABOD is observed as the absolute worst method for outlier treatment with regards to clustering.

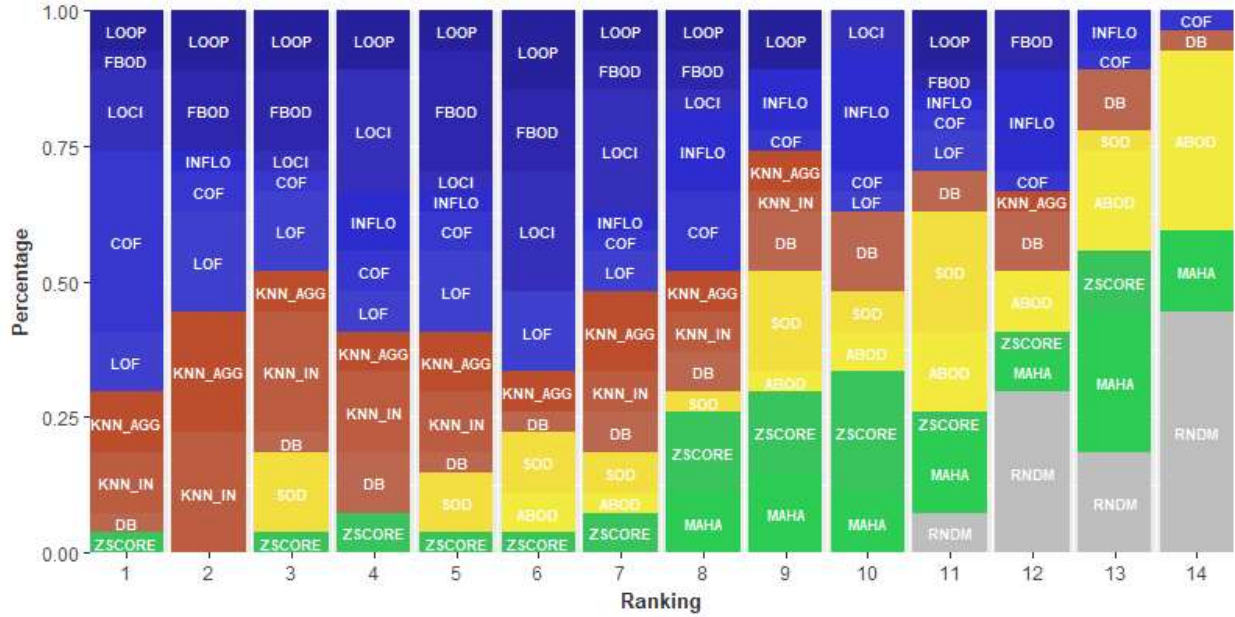


Figure 22: Proportion of treated datasets ranking for k-means clustering across all coherent solutions

5.8. UNDERSTANDING COHERENCE AND CLUSTER RESULTS

Inspecting the clustering results for coherence among the evaluation metrics, it is obvious that the solutions differ in coherence. As a great example of the deviating results, *Frogs* and *BreastCancerDiagnostic* datasets have the best and worst coherence, respectively.

Frogs has an improvement of clustering results of almost 10% with hierarchical clustering, given a dataset treated with FBOD (15% outlier removal). This solution has the lowest cluster index, the strongest cluster correlation, the third greatest Dunn, the greatest Silhouette and the lowest Davies-Bouldin, all indicating that the FBOD treated dataset produce the best outcome with respect to clustering objectives.

BreastCancerDiagnostic has an improvement of clustering results around 6% with hierarchical clustering, given a dataset treated with COF (15% outlier removal). This solution has the fifth lowest cluster index, the seventh strongest cluster correlation, the fifth greatest Dunn, the sixth greatest Silhouette and the seventh lowest Davies-Bouldin. In fact, looking apart from R^2 , the greatest evaluation metrics outcomes

are mainly observed for RNDM, MAHA, ABOD and INFLO, all outlier removal methods poorly performing with respect to R^2 .

Thus, with strong confidence, we know that the cluster solution for *Frogs* is the best with respect to minimizing intra-cluster distance and maximizing inter-cluster distance. On the other hand, we have little or no confidence that the cluster solution for *BreastCancerDiagnostic* is the best with respect to minimizing intra-cluster distance and maximizing inter-cluster distance. Consequently, we cannot conclude how or if outlier detection can improve clustering results for *BreastCancerDiagnostics* with contradicting results.

Returning to the quote by Jain & Dubes (1988), the results differing in coherence for evaluation metrics for several datasets (e.g. *BreastCancerDiagnostics* and *ForestFires*) remain unexplained and gives breeding ground to the belief that evaluating clustering results is 'black art'. However, to understand why evaluation metrics have great coherence for some datasets while other datasets returns poor results, it is necessary to understand to which extent data actually contains clusters and how data is structured. The issue of understanding the structure of data, consist of dealing with a great number of dimensions, making it impossible to inspect visually. Moreover, Hopkins statistic has no relationship with coherence. One could argue that the simple solution would be to run an arbitrary clustering algorithm and inspect the R^2 for rapid increases, indicating if large clusters – or subclusters – are present in data. However, it does not appear to be an appropriate method when comparing the clustering results of *Ionosphere* and *ForestFires*, having coherent and non-coherent results, respectively.

To understand the structure of data, dimension reduction is useful. The purpose of dimension reduction is to express as much information as possible with as few dimensions as possible. For dimension reduction, principal component analysis comes to mind. However, being a linear method possess a weakness as we cannot assume linearity in data space. Thus, a non-linear method of dimensionality reduction is more appropriate.

For non-linear dimension reduction, van der Maaten and Hinton (2008) suggested an algorithm called t-distributed stochastic neighbor embedding (t-SNE). The algorithm mix all observations randomly in a two-dimensional space such that two uniform distributions are created. With a distance matrix as input, the distribution of distances for each observation x_i , to every other observation x_j , is computed based on a normal distribution $N(0,1)$. Logically, distant observations to x_i will end up in the tails with a low probability while close observations will center around x_i with a probability of 0.4. A new matrix for

observations are computed, being the probability for x_i to have x_j as a neighbor. Given a certain number of iterations, the algorithm iterates over every observation, and based on probability matrix, observation x_i is attracted to other observations of which a high probability of being a neighbor is present, and repelled by observations of which a low probability of being a neighbor is present.

For the two datasets *Frogs* and *BreastCancerDiagnostic*, the t-SNE dimension reduction with a perplexity of 30 (not described above) and 1000 iterations, is shown in Figure 23. The results of the clustering makes greater sense when taking the structure of the data in to account.

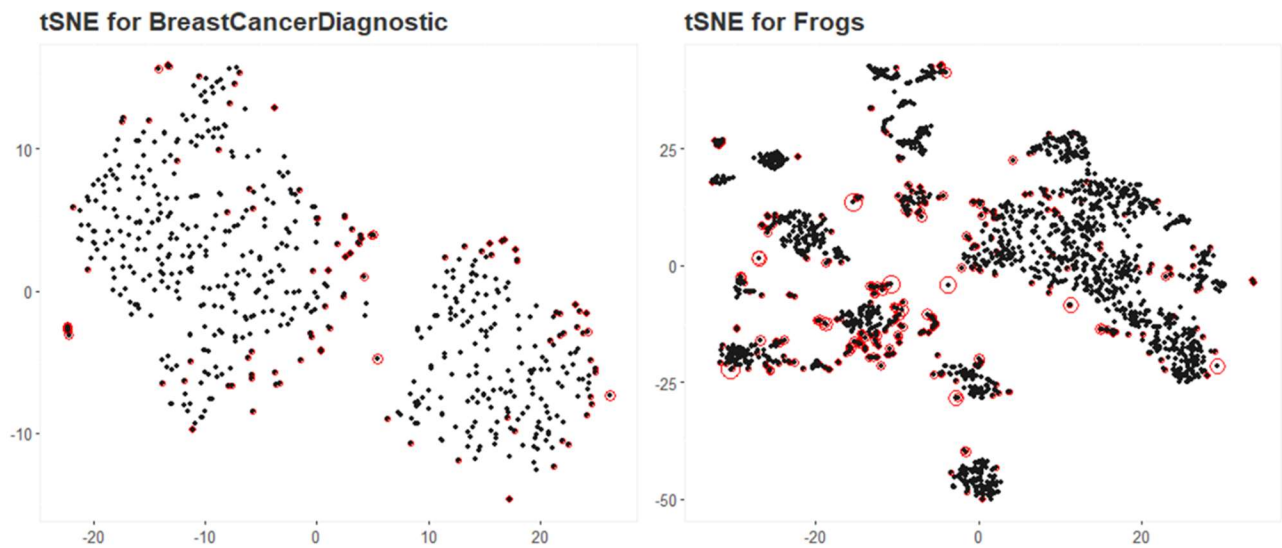


Figure 23: tSNE applied *BreastCancerDiagnostic* (left) and *Frogs* (right) with LOF applied to original data and projected

For *BreastCancerDiagnostics*, the UCI Machine Learning Repository information provided tells that the single categorical variable is binary and classifies breast cancer tumors as malignant or benign tumors. The binary variable therefore divides the dataset in to two major clusters, explaining the rapid increase for R^2 , when going from one to two clusters (see: annex). Applying any arbitrary outlier method to *BreastCancerDiagnostics* removes the observations on the edge of the gaussian clusters, making the two major clusters more dense. Applying more than two clusters to *BreastCancerDiagnostics* cause the two clusters to split in to several subclusters, although they are not clearly separated in data. Thus, R^2 will increase and more variance will appear explained, but Silhouette, cluster correlation and cluster index will not improve, explaining why R^2 for *BreastCancerDiagnostic* has an incoherence for evaluation metrics in Table 7.

Inspecting Table 18, the problem of coherence is apparent in the *BreastCancerDiagnostics* dataset. For the RNDM treated data, an R^2 of 0.28 is observed with a solution of two clusters compared to 0.39 for LOF treated data, while the cluster index is 0.66 compared to 0.59 for LOF treated data. With a solution of five clusters, R^2 increase to 0.47 for RNDM treated data compared to 0.52 for LOF treated data. However, the cluster index decrease to 0.54 for RNDM treated data compared to an increase to 0.65 for LOF treated data. Thus, for a solution of five clusters, the ratio between intra- and inter-cluster distance improves for RNDM treated data as true outliers are not removed, while LOF has a worsened ratio between intra- and inter-cluster distance. For a solution of two clusters, the LOF treated data has the greatest R^2 due to removed outliers but also has the best cluster index of 0.59.

Clusters	1	2	3	4	5	6	7	8	9	10
LOF – R^2	0	0.39	0.45	0.49	0.52	0.55	0.58	0.60	0.61	0.62
RNDM – R^2	0	0.28	0.37	0.43	0.47	0.50	0.52	0.54	0.56	0.58
LOF – cls. indx.	NA	0.59	0.58	0.63	0.65	0.61	0.59	0.59	0.59	0.58
RNDM – cls. indx	NA	0.66	0.59	0.55	0.54	0.53	0.53	0.56	0.55	0.56

Table 18: Results for hierarchical clustering for the *BreastCancerDiagnostics* dataset

For *Frogs*, the UCI Machine Learning Repository information provided tells that ‘dataset was created segmenting 60 audio records belonging to 4 different [Frog] families, 8 genus, and 10 species’. Naturally, many clusters are present as frog species have different characteristics and inspecting the right plot in Figure 23, some observations are outliers between the clusters. The outlier algorithm LOF successfully identifies these outlying observations for removal, increasing the inter-cluster distance. Contrary to *BreastCancerDiagnostics*, applying more than two clusters does not possess a problem as data is clearly separable. Thus, when R^2 increase Silhouette, cluster correlation and cluster index will improve, explaining why R^2 for *Frogs* has a great coherence for evaluation metrics in Table 7.

Inspecting Table 19, an R^2 of 0.33 is observed with a solution of two clusters for RNDM treated data compared to 0.37 for LOF treated data, while the cluster index is 0.65 for RNDM treated data compared to 0.59 for LOF treated data. With a solution of five clusters, R^2 increase to 0.54 for RNDM treated data compared to 0.62 for LOF treated data, while the cluster index decrease to 0.51 for RNDM treated data compared to 0.46 for LOF treated data. The correlation between R^2 and cluster index is -0.98. Thus, LOF treated data returns better results regardless of the number of clusters, reflecting the characteristics of the data.

Clusters	1	2	3	4	5	6	7	8	9	10
LOF – R^2	0	0.37	0.48	0.55	0.62	0.67	0.71	0.74	0.76	0.78
RNDM – R^2	0	0.33	0.43	0.49	0.54	0.58	0.62	0.65	0.67	0.69
LOF – cls. indx.	NA	0.59	0.51	0.48	0.46	0.45	0.43	0.41	0.40	0.39
RNDM – cls. indx	NA	0.65	0.55	0.53	0.51	0.51	0.48	0.47	0.45	0.43

Table 19: Clustering results for hierarchical clustering for the Frogs dataset

Based on the differences of cluster results for the two datasets *BreastCancerDiagnostics* and *Frogs*, it is concluded that evaluation metrics Davies-Bouldin, Silhouette, Dunn, cluster correlation and cluster index are indicators for users to know how many clusters to choose as a solution, whereas R^2 is a tool to describe explained variance.

To exemplify the differences in clustering results for *Frogs* and *BreastCancerDiagnostics*, Figure 24 is created. It should be noticed that clustering is applied to tSNE output and not original data, explaining the perfect divide in clusters. Moreover, LOF is applied and 15% of observations are removed. Inspecting left plot in Figure 24, *BreastCancerDiagnostics* only has great inter-cluster distance with a cluster solution of two clusters. Increasing the number of cluster cause the two major clusters to split into subclusters resulting in low inter-cluster distance. For right plot *Frogs* has several natural clusters with great inter-cluster distance. Increasing the number of clusters cause the small clusters to split in to subclusters, still having great inter-cluster distance. The exception is only the major cluster in the right side, splitted into four subclusters with low inter-cluster distance.

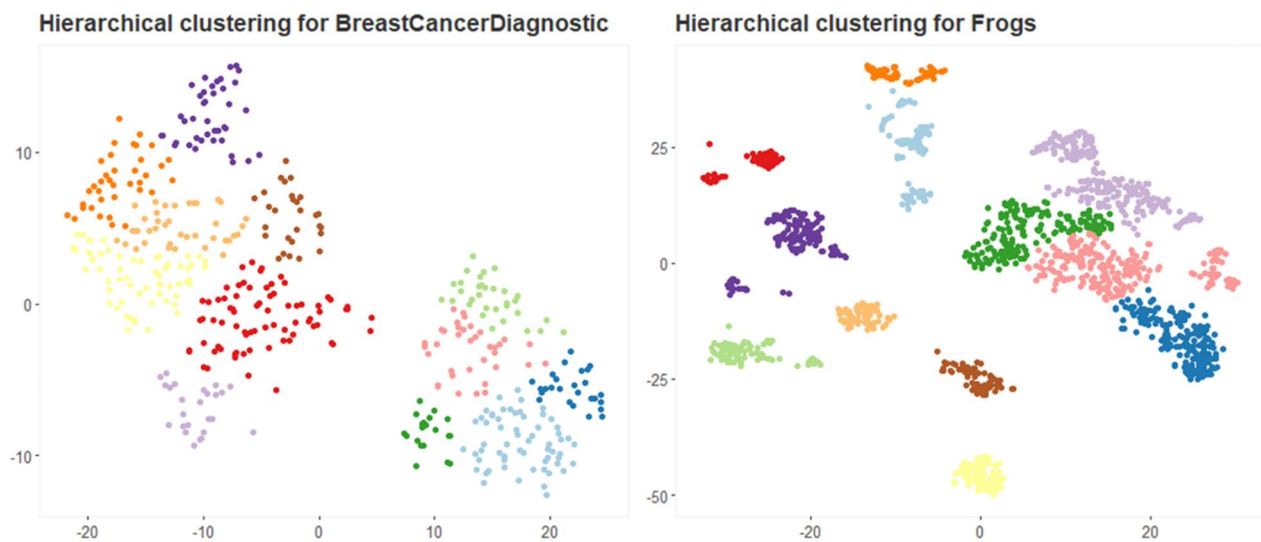


Figure 24: Hierarchical clustering for *BreastCancerDiagnostics* (left) and *Frogs* (right) with 12 clusters

5.9. UNDERSTANDING REMOVAL PERCENTAGE

When carrying out cluster analysis in real life, outlier detection should be a case from dataset to dataset and only true outliers should be removed instead of an arbitrary percentage of observations. In addition to previous discussion about coherence for datasets, it is also observed that the treated dataset have a different increase in R^2 when additional 5% of observations are removed. For the analysis in this thesis, we removed 5%, 10% and 15% of outliers, to have an equal comparison of across datasets and algorithms. The outcome is however different from dataset to dataset. A great example of the different outcomes, is the case of outliers for *HeartDisease* and *Ionosphere*. *HeartDisease* has an average increase in R^2 of 0.01 for removal of additional 5% outliers, while *HeartDisease* has an average increase of 0.053.

Inspecting Table 20, *Ionosphere* greatest outlying observation has an LOF score of 5.89, indicating that neighboring observations in average have 5.89 times more dense neighborhoods, compared to 1.71 for *HeartDisease*. For the 15th percentile observation for *Ionosphere* an LOF of 2.37 is observed, in comparison to 1.14 for *HeartDisease*. Thus, it is clear that *Ionosphere* has fractured data characteristics with many outliers while *HeartDisease* has dense data characters with few outliers. One could argue that it would be correct to remove just 1% of observations for *HeartDisease* while a discussion of removal percentage for *Ionosphere* is more difficult.

Outlier percentile	0%	2.5%	5%	7.5%	10%	12.5%	15%
Ionosphere	5.89	4.17	3.18	2.81	2.61	2.45	2.37
HeartDisease	1.71	1.25	1.20	1.17	1.16	1.15	1.14

Table 20: LOF scores for different percentiles for *Ionosphere* and *HeartDisease* datasets given $k=30$

5.10. UNDERSTANDING ADVANTAGES AND DISADVANTAGES OF OUTLIER METHODS

Inspecting the results for SSW and R^2 , it is observed that DB and Z-score are great at reducing variance but often differ from each other in ability to explain variance. It is also observed that the density-based approaches LOF, COF, LOCI, LOOP and the distance-based approaches KNN_IN and KNN_AGG frequently return the best results. These findings are demonstrated in Figure 25.

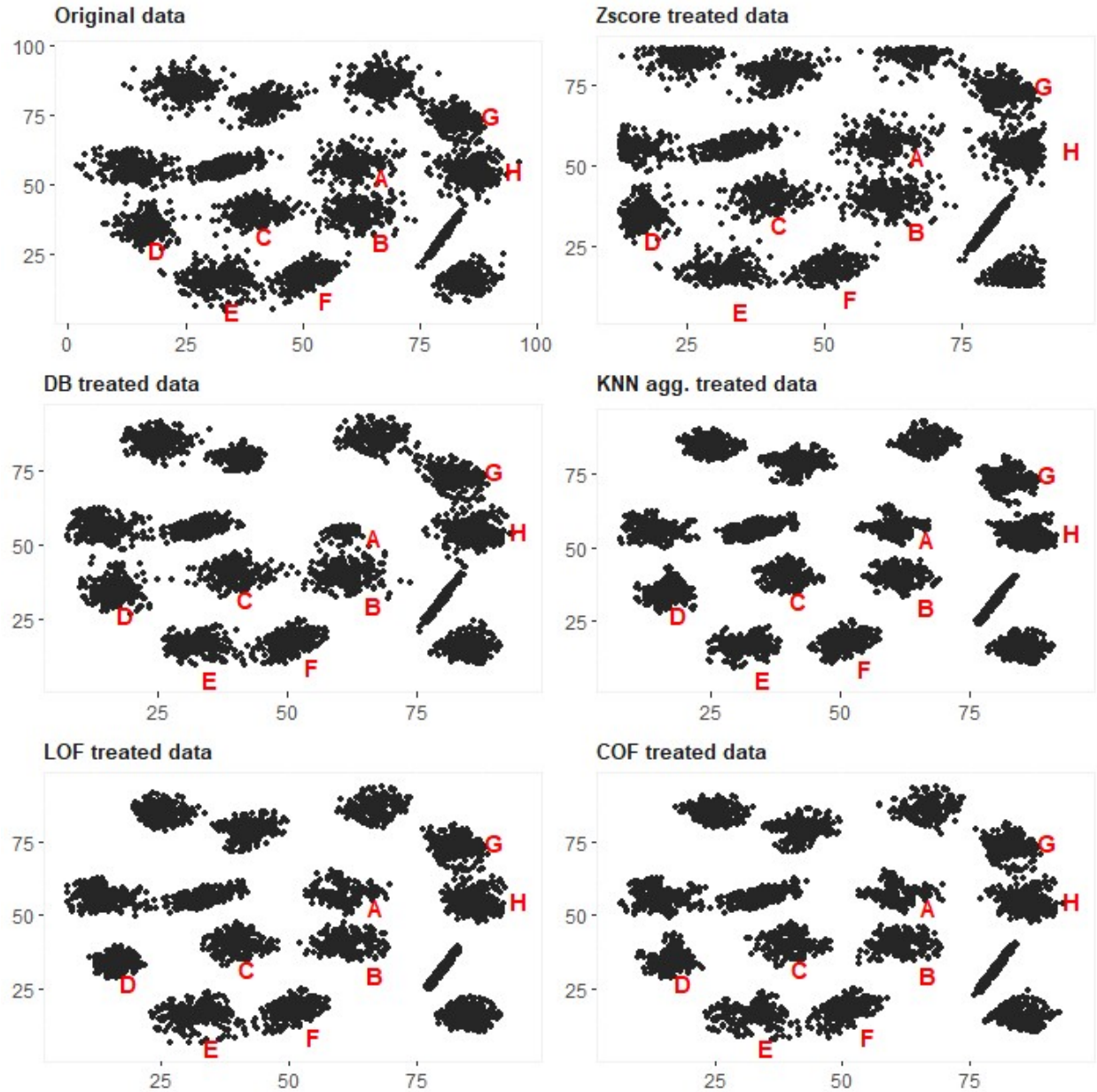


Figure 25: Various outlier detection methods applied to synthetic data

The top left plot in Figure 25 contains original data with 15 clusters having some outliers in between clusters and some outliers deviating from the gaussian clusters at the border of data space. Cluster A and B and G and H are in particular poorly separated. Top right plot has Z-score applied, mid right has kNN aggregated applied, mid left has DB applied, bottom left has LOF applied and bottom right has COF applied. COF, LOF and kNN have setting $k = 30$. All treated datasets have 10% of outlying observations removed.

For Z-score, the result is removal of the observations on the border of data space and halving of several clusters, leaving the outliers in between the clusters intact. For DB, the result is removal of observations on the border of data space and some in between the clusters. DB fails to separate the clusters of A and B, C and D, G and H. Furthermore, half of cluster A is removed such that only the lower half is left. For kNN aggregated, the result is removal of observations on the border of data space and in between the clusters. Separation of cluster A and B, C and D, E and F is clear. For both LOF and COF, the result is removal of observation on the border of data space and in between the clusters. However, cluster E and F and G and H are not completely separated.

The weakness of Z-score is here demonstrated. For most datasets in analysis, the best results for Z-score are observed when only 5% of observations are removed, indicating that few extreme outliers are removed in the border of data space, returning better clustering results compared to when 10% or 15% of observations are removed. Z-score is therefore a great tool to reduce overall variance by simply removing outliers on the border of data space, but fails to improve inter-cluster distance. Thus, Z-score only has justification when users are working with data structures characterized by one major cluster or multiple clusters only with outliers on the border of data space.

The weakness of DB is demonstrated as being a 'global' method with a spherical neighborhood. With its global comparison of density in neighborhood, to density of neighborhood for every other observation in data, the problem occurs for observations at the border of data space. When applied to observations at the border of data space, the neighborhood sphere exceeds the actual data space, meaning a smaller proportion of the neighborhood sphere is within the actual data space identifying neighbors. The issue here is described in the introduction of density-based approaches, as the global comparison issue.

Density-based and distance-based approaches LOF, COF and kNN are demonstrated with greater properties to deal with issues of outliers in between the clusters. These approaches therefore have greater justification when working with complex clustering containing multiple 'real' clusters in data. The greatest disadvantage of these approaches is the question of choosing an appropriate k , for comparison of neighborhoods/distances.

Overall evaluation metrics for synthetic data treated with outlier methods are shown in the annex.

6. CONCLUSION

This thesis aimed to contribute to the knowledge on how outlier detection can improve clustering results for the two algorithms k-means and hierarchical clustering. To understand how outlier detection can influence clustering results, a literature review was conducted with the aim of explaining the essentials and identifying tools of outlier detection and clustering. The identified tools were included in an analysis of 9 real life datasets where 11 outlier algorithms and 2 statistical outlier methods were applied and returned treated datasets. The treated datasets had k-means and hierarchical clustering applied. For applied clustering ambiguous results showed that no single outlier algorithm performs best across all datasets, but distance-based approaches KNN_IN and KNN_AGG and density-based approaches LOF, LOCI, LOOP, FBOD and COF generally returned the best results. It was discovered that COF treated datasets have a great performance with k-means applied, compared to the same COF treated datasets for hierarchical clustering.

The general results demonstrated that outlier detection does indeed improve clustering results, but it requires great effort to get the optimal results and can only be achieved to an extent where data allows it. With synthetic data, it was demonstrated that data containing multiple clusters returns the best clustering results when data is treated with distance- or density-based approaches. These two approaches excel at minimizing intra-cluster distance and maximizing inter-cluster distance. For data containing one or two close clusters it was found that Z-score, DB or kNN excel at creating denser clusters and minimize intra-cluster distance but naturally cannot increase inter-cluster distance.

Thus, the issue of knowing when some algorithms excel while others do not was discovered to be a matter of characteristics of data, and in particular knowing if data contains 'real' separated clusters. The multidimensional nature of clustering problems makes it difficult to explore data with summary statistics, leaving users in the unknown with no choice but to apply all algorithms and inspect outcome of evaluation metrics. It was found that dimension reduction algorithms, such as tSNE, can provide guidance for users to avoid applying all methods but that the literature still needs further development to overcome this issue. No relationship between Hopkins statistics and the number of real clusters was found, leading to the conclusion that Hopkins statistic express overall cluster tendency but does not work for indicating a certain number of clusters in data.

Interpreting clustering results with evaluating metrics was also found to be a difficult discipline. The literature lacks a unified approach for clustering evaluation, explaining why methods such as the elbow

method is still around. The identified evaluation metrics were found to have different properties to describe the ratio between intra-cluster and inter-cluster distance.

It was found that R^2 is a tool indicating to which extent variance is explained rather than a tool useful to determine the achievement of clustering objectives. The other evaluation metrics, Davies-Bouldin, Dunn, Silhouette and cluster correlation were found to be useful to identify how many clusters users should choose for a cluster solution. For Davies-Bouldin, Dunn and Silhouette it was found that they have a bias towards the worst structure of clusters in data and that both Dunn and Davies-Bouldin lack properties to deal with clusters of different sizes. To overcome the issue of missing robustness, a new evaluation metric called *cluster index* was introduced as a simple but robust method reflecting the real ratio of intra-cluster and inter-cluster distance.

When applying clustering to outlier treated datasets, datasets with several clearly separated clusters were found to have a linear relationship for evaluation metrics, while datasets with few not clearly separated clusters had no relationship for evaluation metrics. This led to the division of datasets in *coherent* and *incoherent* datasets. For coherent datasets, *HeartDisease*, *PimaIndians*, *Frogs*, *Ionosphere* and *WineQuality*, a linear dependency between cluster evaluation metrics was observed, meaning R^2 is an indicator for achieving cluster objectives. K-means returned the best clustering results for three of five datasets and hierarchical clustering returned the best clustering results for two of five datasets. The greatest improvement of clustering results was observed with hierarchical clustering for four out of five datasets compared to one out of five datasets with k-means. The greatest improvement was furthermore achieved two times with COF, two times with KNN_AGG and one time with FBOD.

For the incoherent datasets, *BreastCancerOriginal*, *ForestFires*, *BreastCancerDiagnostics* and *GesturePhase*, it was difficult to determine which outlier and clustering algorithms performed best. Although a great increase in R^2 was observed, the increased ability to explain variance was not coherent with clustering objectives.

7. RECOMMENDATIONS FOR FUTURE RESEARCH

7.1. CONNECTING OUTLIER DETECTION AND CLUSTERING

Initially, when looking for literature describing how outlier detection can improve clustering results, key phrases such as '*data preparation cluster analysis*', '*outliers data preparation*', '*outliers modelling*', '*outlier clustering*' and '*outliers CRISP-DM*' were used on ScienceDirect, ResearchGate and various search engines. It was found that a surprisingly little amount of literature deals with the issue of removing outliers prior to applying clustering. The identified literature, mostly being books about data preparation (Aggarwal, 2015; Pyle, 1999), approach outlier detection as observations either incorrectly generated or being in the tails of distributions. That is, if observations are considered correctly generated, outlier detection is a statistical approach.

With respect to the findings in this thesis, the literature lacks focus on how to utilize the theory provided to remove outlying observations between the clusters and improve results for clustering objectives. Future research should be concerned with this topic.

7.2. CONNECTING DIMENSION REDUCTION AND CLUSTERING

Initially, when clustering results returned contradicting outcomes, frustration was largely that data could not be inspected visualized multi-dimensionally and that no metrics were available for assisting a decision of an outlier removal percentage. With no metrics or visual methods available, it was decided to remove top 5%, 10% and 15% of observations across all datasets for an equal comparison. This led to missing coherence for cluster evaluation metrics. After a great time was spent contemplating about the incoherent results, dimension reduction was discovered as a great tool to provide an insight in to the data structures of the chosen datasets. Among the dimension reduction tools, Principal Component Analysis (PCA), Self-Organizing Maps (SOM), Multidimensional Scalling (MDS), Sammon Mapping, and tSNE were applied. tSNE showed the best results when comparing it to the output of clustering algorithms. Thus, dimension reduction is a great field of study that allows users to understand data structures. Future research should be concerned with this topic.

8. REFERENCES

- Peirce, B. (1852). Criterion for the Rejection of Doubtful Observations. *Astronomical Journal*. Vol. 2, iss. 45, pp. 161-163
- Pierce, C. (1873). Report of the Superintendent of the U.S. Coast Survey Showing the Progress of the Survey During the Year 1870.
- Driver, H. E., and Kroeber (1932), A. L. Quantitative expression of cultural relationships. *University of California Publications in Archaeology and Ethnology*, 1932, No. 31, pp. 211–216.
- Mahalanobis, P. (1936). On The Generalized Distance in Statistics. *Proceedings of the National Institute of Sciences of India*. Vol. 2, No. 1, pp. 49-55.
- Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. *The Annals of Mathematical Statistics*. Vol. 21, No. 1, 1950, pp. 27-58
- Hopkins, B. (1954). A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany*. Vol. 18, pp. 213–227
- Wilk, M., & Gnanadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. *Biometrika*. Vol. 55, pp. 1-18
- Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics*. Vol. 11, pp. 1-21
- Tobler, W. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*. Vol. 46, pp. 234-240
- Kettenring, J., & Gnanadesikan, R. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics*. Vol. 28, pp. 81-124
- Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*. Vol 4, pp. 95–104
- Tukey, J. (1977). *Explorative Data Analysis*. Pearson. ISBN: 978-0201076165
- Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*. Vol. 19, No. 1., pp. 15-18
- Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 1, No. 2, pp. 224-227
- Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall. ISBN: 978-94-015-3994-4
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. Vol. 20, pp. 53-65

- Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice Hall. ISBN: 978-01-302-2278-7
- Garrett, R. G. (1989). The chi-square plot: a tool for multivariate outlier recognition. Journal of Geochemical Exploration. Vol. 32, pp. 319-341
- Iglewicz, B., & Hoaglin, D. (1993). How To Detect and Handle Outliers. ASQC Quality Press. ISBN: 9780873892476
- Barnett, V., & Lewis, T. (1994). Outliers in statistical data. Wiley. ISBN: 0-471-93094-6
- Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in. AI Magazine, Vol. 17, No. 3, pp. 37-54
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In KDD-96 Proceedings, Vol. 96, No. 34, pp. 226-231
- Arning, A., Agrawal, R., & Raghavan, P. (1996). A Linear Method for Deviation in Large Databases. In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), pp. 164-169
- Knorr, M., & Ng, R. T. (1997). A Unified Approach for Mining Outliers. In Conf. of the Centre for Advanced Studies on Collaborative Research (CASCON). Toronto, Canada. pp. 236-248
- Knorr, M., & Ng, R. T. (1998). Algorithms for Mining Distance-Based Outliers in Large Datasets. In International Conference on Very Large Data Bases. pp. 392-403
- Knorr, M., & Ng, R. T. (1999). Finding Intensional Knowledge of Distance-Based Outliers. In International Conference on Very Large Data Bases. pp. 211-222
- Pyle, D. (1999). Data Preparation for Data Mining. Morgan Kaufmann. ISBN: 978-15-586-0529-9
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, Vol. 5, No. 4, pp. 13-22
- Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. In International Conference on Management of Data. Dallas, TX. pp. 427-438
- Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF : Identifying Density-Based Local Outliers. In Int. Conf. On Management of Data. Dallas, TX. pp. 93-104
- Maesschalck, R. De, & Massart, D. L. (2000). The Mahalanobis distance. Chemometrics and Intelligent Laboratory Systems 50. Vol. 50, No. 1., pp. 1-18
- Jin, W., Tung, A., & Han, J. (2001). Mining Top-n Local Outliers in Large Databases. In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD). San Francisco, CA. pp. 293-298
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Information Systems. Vol. 17, No. 2-3, pp. 107-145

- Tang, J., Chen, Z., Fu, A. W. C., & Cheung, D. W. (2002). Enhancing Effectiveness of Outlier Detections for Low Density Patterns. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). Taipei. pp. 535-548
- Angiulli, F., & Pizzuti, C. (2002). Fast Outlier Detection in High Dimensional Spaces. In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD). Helsinki, Finland. pp. 15-26
- Papadimitriou, S., Gibbons, P. B., & Faloutsos, C. (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral. In International Conference on Data Engineering. pp. 315-326
- Hautamaki, V., & Ismo, K. (2004). Outlier Detection Using k-Nearest Neighbour Graph. In International Conference on Pattern Recognition. Cambridge, UK. pp. 430-433
- Lazarevic, A., & Kumar, V. (2005). Feature Bagging for Outlier Detection. In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD). Chicago, IL. pp. 157-166
- Filzmoser, P., Garrett, R. G. and Reimann, C. (2005). Multivariate Outlier Detection in Exploration Geochemistry. Computers & Geosciences. Vol. 31, No. 5, pp. 579-587
- Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2005). Introduction to Data Mining (Second edition). ISBN: 978-03-213-2136-7
- SAS. (2006). SEMMA. http://faculty.smu.edu/tfomby/eco5385_eco6380/data
- Jin, W., Tung, A. K. H., Han, J., & Wang, W. (2006). Ranking Outliers Using Symmetric Neighborhood Relationship. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). Singapore. pp 577-593
- Fan, H., Zaïane, O. R., Foss, A., & Wu, J. (2006). A Nonparametric Outlier Detection for Effectively Discovering Top-N Outliers from Engineering Data. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). Singapore. pp. 557-566
- Filzmoser, P., Maronna, R. G., & Werner, M. (2006). Outlier Identification in High Dimensions. Computational Statistics & Data Analysis. Vol. 52, No. 3, pp. 1694-1711
- Latecki, L., Lazarevic, A., & Pokrajac, D. (2007). Outlier Detection with Kernel Density Functions. In MLDM: International Workshop on Machine Learning and Data Mining in Pattern Recognition. pp. 61-75
- Vanderviere, E., & Huber, M. (2008). An Adjusted Boxplot for Skewed Distributions. Computational Statistics & Data Analysis. Vol. 52, No. 12, pp. 5186-5201
- Kriegel, H.-P., Schubert, M., & Zimek, A. (2008). Angle-Based Outlier Detection in High-dimensional Data. In Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD). Las Vegas, NV. pp. 444-452
- Maaten, L. Van Der, & Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research. Vol. 9, pp. 2579-2605

- Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). Bangkok, Thailand. pp. 831-838
- Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2009). LoOP: Local Outlier Probabilities. In ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China. pp. 1649-1652
- Zhang, K., Hutter, M., & Jin, H. (2009). A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). pp. 813-822
- Kriegel, H.-P., Kröger, P., Schubert, E., & Zimek, A. (2011). Interpreting and Unifying Outlier Scores. In International Conference on Data Mining. Arizona. pp. 13-24
- Gao, J., Hu, W., Zhang, Z., Zhang, X., & Wu, O. (2011). RKOF: Robust Kernel-Based Local Outlier Detection. In Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD). pp. 270-283
- Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. Statistical Analysis and Data Mining. pp. 363–387.
- Provost, F., & Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (First edition). O'Reilly Media. ISBN: 978-14-493-6132-7
- Aggarwal, C. C., & Reddy, C. K. (2013). Data Clustering: Algorithms and Applications. Chapman and Hall. ISBN: 978-14-665-5821-2
- Schubert, E., Arthur, Z., & Kriegel, H.-P. (2014). Generalized Outlier Detection with Flexible Kernel Density Estimates. In SIAM International Conference on Data Mining. pp. 542-550
- Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer. ISBN: 978-3-319-14141-1
- Tang, B., & Haibo, He. (2017). A local density-based approach for outlier detection. Neurocomputing. Vol. 241, pp. 171-180

9. ANNEX

9.1. SIMILARITY MATRICES FOR DATASETS

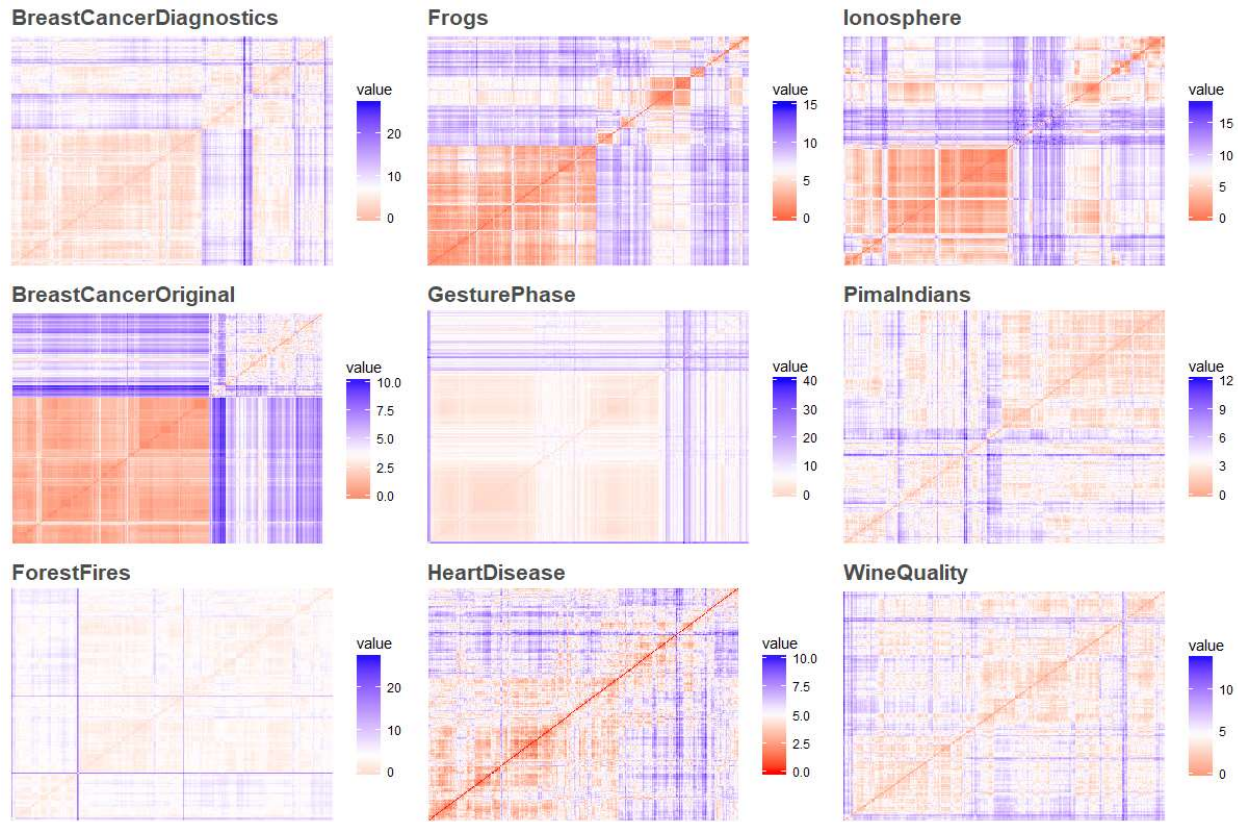


Figure 26: Similarity matrices for datasets chosen for outlier treatment

Comment: The similarity matrices are ordered by rownumber. For each similarity matrix a low distance is indicated by red colour and high distance is indicated by blue colour. Clusters are present when a great share of distances are visualized in red.

9.2. EVALUATION METRICS FOR SYNTHETIC DATA TREATED WITH OUTLIER METHODS

	Davies-Bouldin	Silhou	Dunn	ClusterCor	ClusterIndex	R ²
Zscore	0.430	0.723	0.046	-0.496	0.106	0.985
DB	0.419	0.726	0.043	-0.509	0.098	0.987
KNN agg.	0.378	0.761	0.101	-0.502	0.092	0.989
LOF	0.406	0.742	0.053	-0.501	0.097	0.988
COF	0.414	0.738	0.068	-0.499	0.099	0.987

Table 21: Evaluation metrics applied to dataset with hierarchical cluster solution with the data treated by the respective outlier methods

Comment: The evaluation metrics has small differences due to the strong gaussian tendency for the majority of the clusters.

9.3. COMMENTS ON RESULTS FOR INCOHERENT DATASETS

For the tables displaying results for R^2 , the colourization of methods are according to their respective outlier detection approach. Distance-based approaches KNN_AGG, KNN_IN and DB are red. Density-based approaches LOCI, LOOP, LOF, COF, INFLO and FBOD are blue. FBOD is considered a density-based approach, as the input algorithms used for the algorithm itself are density-based. Statistical approaches MAHA and ZSCORE are green. High-dimensional approaches ABOD and SOD are yellow and random removed observations (RNDM) is grey. For each column in method, the ranking is based on the result of R^2 for the respective removal percentage (5%, 10% and 15).

For demonstration purpose, see table 22. The greatest results for *BreastCancerDiagnostics* with respect to R^2 with hierarchical clustering applied with 5% outlier removal, is achieved with LOCI with an R^2 of 0.761. With 15% outlier removal, the greatest result with respect to R^2 , is achieved with KNN_AGG with an R^2 of 0.785.

9.3.1. BreastCancerOriginal

For hierarchical clustering the greatest improvement of clustering results is achieved with the KNN_AGG treated dataset (15%), returning an R^2 of 0.785, 0.046 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, LOOP, INFLO and LOCI, the distance-based approaches KNN_IN and KNN_AGG and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.009 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD, the distance-based method DB and the density-based methods COF and FBOD. Of these solutions, DB and ABOD returns worse results when compared to RNDM treated dataset. For these solutions, R^2 decrease with 0.001 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 863 (15%) followed by ZSCORE and DB with an SSW of 920 and 970, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.785, 0.743 and 0.707, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOCI	LOCI	KNN_AGG	0.761	0.761	0.785	1620	1620	863
2	LOF	LOF	KNN_IN	0.758	0.758	0.781	1640	1640	1289
3	KNN_IN	KNN_IN	SOD	0.757	0.757	0.777	1607	1607	1145
4	KNN_AGG	KNN_AGG	LOCI	0.753	0.753	0.768	1433	1433	1499
5	LOOP	LOOP	INFLO	0.747	0.747	0.768	1714	1714	1123
6	INFLO	INFLO	LOF	0.744	0.744	0.766	1552	1552	1519
7	SOD	SOD	LOOP	0.743	0.743	0.755	1602	1602	1577
8	MAHA	MAHA	MAHA	0.742	0.742	0.749	1665	1665	1375
9	COF	COF	COF	0.740	0.740	0.747	1762	1762	1634
10	RNDM	ZSCORE	ZSCORE	0.739	0.737	0.743	1754	1441	920
11	ZSCORE	FBOD	RNDM	0.737	0.737	0.739	1441	1781	1479
12	FBOD	RNDM	FBOD	0.733	0.733	0.736	1809	1772	1698
13	DB	DB	ABOD	0.732	0.732	0.709	1443	1443	1243
14	ABOD	ABOD	DB	0.722	0.722	0.707	1518	1518	970

Table 22: Hierarchical clustering results for BreastCancerOriginal

For k-means clustering the greatest improvement of clustering results is achieved with KNN_IN treated dataset (15%), returning an R^2 of 0.783, 0.057 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, INFLO, LOCI and LOOP, the distance-based approaches KNN_IN and KNN_AGG and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.009 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD, the distance-based method DB and the density-based methods COF and FBOD. Of these solutions, DB, ABOD and FBOD returns worse results when compared to RNDM treated dataset. For these solutions, R^2 decrease with 0.001 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 888 (15%) followed by ZSCORE and DB with an SSW of 918 and 949, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.783, 0.744 and 0.714, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOF	KNN_AGG	KNN_IN	0.748	0.768	0.783	1705	1121	1279
2	KNN_AGG	INFLO	KNN_AGG	0.746	0.757	0.779	1468	1312	888
3	KNN_IN	KNN_IN	SOD	0.746	0.757	0.767	1682	1545	1195
4	LOCI	LOCI	INFLO	0.746	0.754	0.760	1722	1627	1160
5	INFLO	SOD	LOF	0.745	0.751	0.755	1545	1418	1586
6	LOOP	LOF	LOOP	0.744	0.751	0.753	1732	1652	1593
7	COF	LOOP	LOCI	0.741	0.746	0.751	1760	1675	1608
8	MAHA	MAHA	COF	0.740	0.745	0.747	1672	1528	1640
9	SOD	COF	MAHA	0.740	0.743	0.744	1617	1703	1401
10	ZSCORE	ZSCORE	ZSCORE	0.738	0.737	0.744	1435	1168	918
11	RNDM	RNDM	FBOD	0.737	0.736	0.732	1738	1678	1711
12	FBOD	FBOD	RNDM	0.736	0.731	0.726	1791	1774	1631
13	DB	DB	DB	0.729	0.730	0.714	1458	1160	949
14	ABOD	ABOD	ABOD	0.720	0.714	0.709	1529	1255	1246

Table 23: K-means clustering results for BreastCancerOriginal

For *BreastCancerOriginal* it is concluded that hierarchical clustering returns the best results, with an R^2 of 0.785 compared to an R^2 of 0.783 for k-means clustering. The margins between the two clustering algorithms are however very small.

When asked how outlier detection can improve clustering results, the greatest impact is observed for k-means clustering, improving the overall clustering results with 0.057, compared to 0.046 for hierarchical

clustering. Interestingly, not all outlier detection methods improve the clustering results. ABOD and DB, being the two worst outlier detection methods for *BreastCancerOriginal*, in average have 0.02 worse R^2 when compared to RNDM, meaning a better outcome is achieved by removing random observations.

Inspecting SSW, it is concluded that KNN_AGG, DB and ZSCORE all excel in creating dense clusters, with the lowest SSW observed. However, they greatly differ in R^2 . As an example, DB has an R^2 around 0.72 for both hierarchical and k-means clustering, while KNN_AGG has an R^2 going from 0.74 to 0.78.

Thus, KNN_AGG, DB and ZSCORE creates the densest clusters in absolute figures but differ in ability to explain the proportion of remaining variance, regardless of clustering methods. That is, being great at minimizing intra-cluster distance and but differing in ability to maximize inter-cluster distance.

For both clustering algorithms, it is concluded that outlier detection explains relatively less variance when removing additional 5% observations.

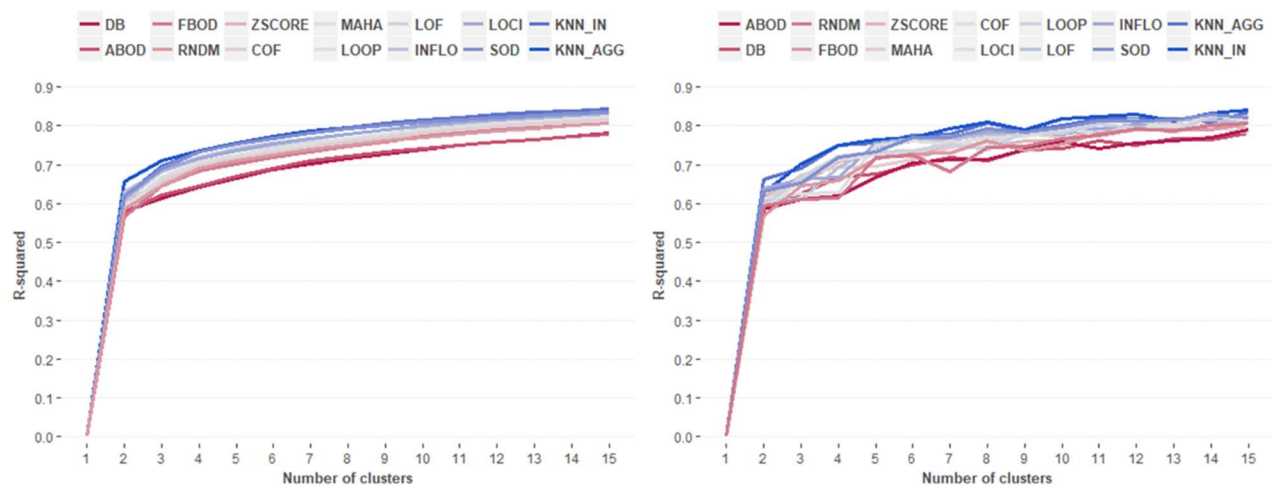


Figure 27: R^2 for *BreastCancerDiagnostic* with hierarchical clustering (left) and k-means (right) with 15% observations removed

9.3.2. ForestFires

For hierarchical clustering the greatest improvement of clustering results is achieved with SOD treated dataset (15%), returning an R^2 of 0.509, 0.016 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, LOCI, FBOD, LOOP and COF, the distance-based approaches KNN_IN and KNN_AGG and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.010 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approach ABOD, the distance-based approaches DB and KNN_IN and the density-based approach INFLO. Of these solutions, only KNN_IN does not returns worse results when compared to RNDM treated dataset. Furthermore, R^2 increase with 0.003 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 1710 (15%) followed by FBOD and LOOP both with an SSW of 1719. The solutions almost have equal ability to explain the remaining variance, with an R^2 just around 0.500.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	SOD	SOD	SOD	0.496	0.509	0.509	2528	2353	2353
2	FBOD	LOF	LOF	0.479	0.502	0.502	2029	1724	1724
3	COF	COF	COF	0.479	0.502	0.502	2047	1753	1753
4	LOF	LOOP	LOOP	0.478	0.501	0.501	2029	1719	1719
5	KNN_AGG	FBOD	FBOD	0.478	0.501	0.501	2013	1719	1719
6	KNN_IN	KNN_AGG	KNN_AGG	0.477	0.499	0.499	2044	1710	1710
7	LOOP	LOCI	LOCI	0.476	0.495	0.495	2050	1764	1764
8	RNDM	KNN_IN	KNN_IN	0.474	0.495	0.495	2650	1752	1752
9	ZSCORE	RNDM	RNDM	0.472	0.493	0.493	2024	2661	2661
10	LOCI	ZSCORE	ZSCORE	0.471	0.476	0.476	2059	1789	1789
11	DB	INFLO	INFLO	0.462	0.465	0.465	2048	2096	2096
12	INFLO	MAHA	MAHA	0.458	0.461	0.461	2293	2088	2088
13	ABOD	DB	DB	0.457	0.457	0.457	2235	1781	1781
14	MAHA	ABOD	ABOD	0.454	0.451	0.451	2255	2116	2116

Table 24: Hierarchical clustering results for ForestFires

For k-means clustering the greatest improvement of clustering results is achieved with COF treated dataset (15%), returning an R^2 of 0.539, 0.054 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, LOCI, FBOD, LOOP and COF and the distance-based approaches KNN_IN and KNN_AGG. For these solutions, R^2 increase with 0.011 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approaches ABOD and SOD, the distance-based approaches DB and KNN_IN and the density-based approach INFLO. Of these solutions, none returns worse results when compared to RNDM treated dataset. Furthermore, R^2 only increase with 0.003 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by LOF with an SSW of 1420 (15%) followed by LOOP and KNN_AGG with an SSW of 1441 and 1442, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 0.536, 0.536 and 0.519, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	FBOD	LOF	COF	0.512	0.527	0.539	1924	1638	1484
2	LOF	FBOD	LOF	0.510	0.524	0.536	1905	1645	1420
3	LOOP	KNN_AGG	LOOP	0.509	0.523	0.536	1919	1628	1441
4	KNN_AGG	LOCI	KNN_IN	0.509	0.522	0.532	1892	1673	1475
5	COF	COF	FBOD	0.509	0.521	0.530	1926	1683	1450
6	LOCI	LOOP	LOCI	0.508	0.520	0.529	1918	1653	1482
7	ZSCORE	KNN_IN	KNN_AGG	0.504	0.519	0.519	1903	1671	1442
8	KNN_IN	ZSCORE	ZSCORE	0.504	0.509	0.512	1939	1677	1466
9	DB	MAHA	SOD	0.499	0.501	0.506	1907	1935	1998
10	MAHA	ABOD	MAHA	0.491	0.491	0.497	2099	1958	1805
11	ABOD	SOD	INFLO	0.488	0.490	0.490	2108	2439	1844
12	INFLO	DB	DB	0.484	0.488	0.489	2183	1677	1454
13	SOD	INFLO	ABOD	0.480	0.481	0.487	2608	2034	1850
14	RNDM	RNDM	RNDM	0.466	0.471	0.485	2917	2718	2600

Table 25: K-means clustering results for ForestFires

For *ForestFires* it is concluded that k-means clustering returns the best results, with an R^2 of 0.539 compared to an R^2 of 0.502 for hierarchical clustering.

When asked how outlier detection can improve clustering results, the greatest impact is observed for k-means clustering, improving the overall clustering results with 0.054 points, compared to 0.016 for hierarchical clustering. Moreover, R^2 increase by 0.011 for k-means clustering, compared to 0.010 for hierarchical clustering, when removing additional 5% observations.

In general, the hierarchical clustering solution is not satisfactory, considering that RNDM treated dataset has greater R^2 than ZSCORE, INFLO, MAHA, DB and ABOD.

Interestingly, SOD returns the best cluster results across any proportion of removed observations for hierarchical clustering, while it returns results below average for k-means clustering.

Inspecting SSW, KNN_AGG, LOOP, FBOD and LOF have the greatest the greatest variance reduction for both cluster solutions, with the density-based approaches having the most satisfactory result for R^2 . Thus, density-based approaches achieve better results for R^2 , when compared to KNN_AGG and therefore have greater ability to explain variance.

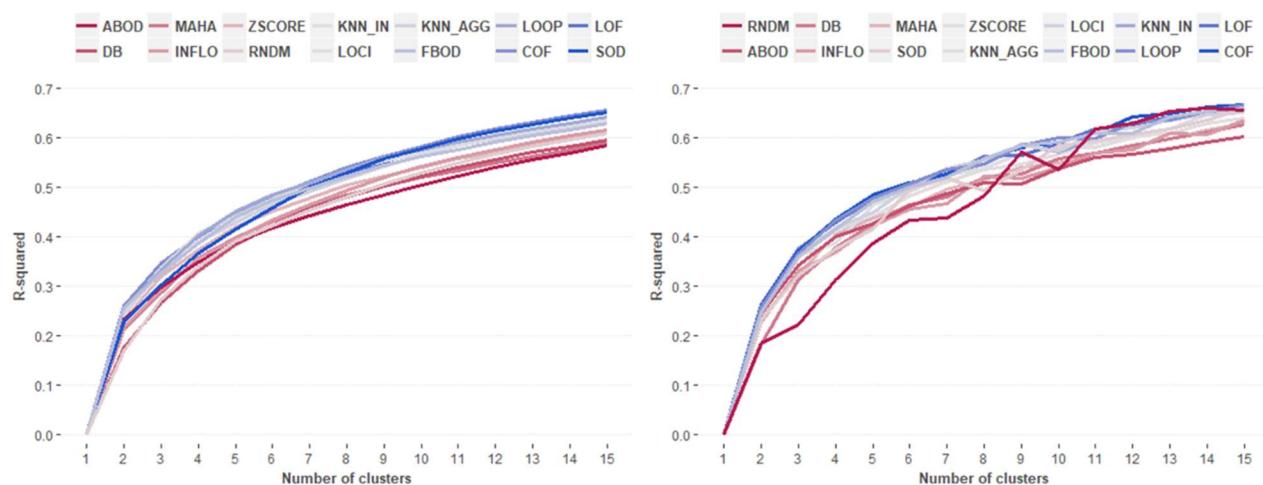


Figure 28: R^2 for ForestFires with hierarchical clustering (left) and k-means (right) with 15% removed observations

9.3.3. BreastCancerDiagnostic

For hierarchical clustering the greatest improvement of clustering results is achieved with the COF treated dataset (15%), returning an R^2 of 0.586, 0.061 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD and LOCI, the distance-based approaches KNN_IN and DB and the statistical test ZSCORE. For these solutions, R^2 increase with 0.012 points in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the distance-based approaches KNN_AGG and DB, the statistical approach MAHA, the high-dimensional approaches SOD and ABOD and the density-based approach INFLO. Of these solutions, SOD, ABOD and MAHA returns worse results than with RNDM treated dataset. For these solutions, R^2 increase with 0.001 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 4298 (15%) followed by KNN_AGG and LOF with an SSW of 4332 and 4411, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.553, 0.557 and 0.579, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	FBOD	LOOP	COF	0.551	0.573	0.586	6284	5305	4520
2	LOOP	LOF	KNN_IN	0.551	0.571	0.579	6403	5231	4793
3	COF	KNN_IN	LOF	0.551	0.571	0.579	6345	5515	4411
4	ZSCORE	COF	LOOP	0.550	0.571	0.574	6284	5208	4556
5	LOF	LOCI	FBOD	0.550	0.565	0.573	6339	5156	4469
6	LOCI	FBOD	LOCI	0.548	0.564	0.571	6379	5158	4479
7	KNN_IN	DB	KNN_AGG	0.542	0.553	0.557	6654	5045	4332
8	KNN_AGG	KNN_AGG	DB	0.541	0.553	0.553	6302	5045	4289
9	DB	ZSCORE	ZSCORE	0.538	0.547	0.541	6381	5143	4454
10	INFLO	INFLO	INFLO	0.537	0.539	0.539	7470	6804	6362
11	MAHA	SOD	SOD	0.524	0.521	0.529	8167	5788	5045
12	RNDM	RNDM	RNDM	0.521	0.521	0.525	8288	7922	7653
13	SOD	MAHA	MAHA	0.519	0.521	0.522	6790	7868	7383
14	ABOD	ABOD	ABOD	0.518	0.516	0.514	7143	6733	6410

Table 26: Hierarchical clustering results for BreastCancerDiagnostic

For k-means clustering the greatest improvement of clustering results is achieved with the COF treated dataset (15%), returning an R^2 of 0.607, 0.059 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches, LOF, COF, FBOD, LOOP and LOCI and the distance-based approaches KNN_IN, KNN_AGG and DB. For these solutions, R^2 increase with 0.003 in average, when removing additional 5% observations.

The seven worst solutions are predominately returned by the the statistical approaches MAHA and ZSCORE, the high-dimensional approaches SOD and ABOD and the density-based approach INFLO. Of these solutions, ABOD and MAHA returns worse results than with RNDM treated dataset. For these solutions, R^2 increase with 0.0001 points in average, when removing additional 5% observations. That is, almost no increase in explained variance, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by DB with an SSW of 4068 (15%) followed by KNN_AGG and FBOD with an SSW of 4069 and 4176, respectively. However, the solutions differ in ability to explain the remaining variance, with an R^2 of 0.577, 0.584 and 0.600, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOOP	LOOP	COF	0.598	0.597	0.607	4989	4999	4288
2	KNN_IN	COF	KNN_IN	0.597	0.597	0.605	5177	4893	4503
3	LOF	KNN_IN	LOOP	0.596	0.597	0.603	4929	5186	4254
4	COF	LOF	LOF	0.596	0.596	0.600	4905	4928	4183
5	FBOD	FBOD	FBOD	0.593	0.593	0.600	4898	4924	4176
6	LOCI	LOCI	LOCI	0.586	0.586	0.597	4901	4906	4213
7	KNN_AGG	DB	KNN_AGG	0.579	0.578	0.584	4751	4764	4069
8	DB	KNN_AGG	DB	0.578	0.575	0.577	4760	4798	4068
9	ZSCORE	ZSCORE	ZSCORE	0.575	0.572	0.568	4831	4866	4191
10	INFLO	INFLO	INFLO	0.562	0.565	0.565	6469	6432	6007
11	SOD	SOD	SOD	0.546	0.546	0.553	5485	5494	4782
12	RNDM	MAHA	RNDM	0.544	0.544	0.548	7463	7479	7086
13	MAHA	ABOD	MAHA	0.544	0.544	0.543	7494	6351	7071
14	ABOD	RNDM	ABOD	0.542	0.539	0.538	6372	7572	6101

Table 27: K-means clustering results for BreastCancerDiagnostic

For *BreastCancerDiagnostic* it is concluded that k-means returns the best clustering results, with an R^2 of 0.607 compared to an R^2 of 0.586 for hierarchical clustering. The margins between the two clustering algorithms are however very small.

When asked how outlier detection can improve clustering results, the greatest impact is observed for hierarchical clustering, improving the overall clustering results with 0.061 compared to 0.059 for k-means. Furthermore, the greatest increase in explained variance is also for hierarchical clustering, with an average increase of 0.012 points, when removing additional 5% observations.

Inspecting SSW, it is concluded that KNN_AGG, DB and ZSCORE all excel in creating dense clusters, with the lowest SSW but fail to remove outliers between clusters, leading to a mediocre R^2 when compared to density-based approaches such as LOOP, LOF and COF. Thus, KNN_AGG, DB and ZSCORE are best at reducing the overall sum of squares (SST) but fail to explain the remaining variance. That is, being great at minimizing intra-cluster distance but failing to maximize inter-cluster distance.

Although being a high-dimensional dataset, SOD and ABOD, both high-dimensional approaches, returns worse results than RNDM, meaning it is a better solution to remove random observations, than remove observations given by the output of SOD and ABOD.

Moreover, among the statistical test, MAHA also produce results worse or equal to RNDM, making it a poor solution for *BreastCancerDiagnostic*.

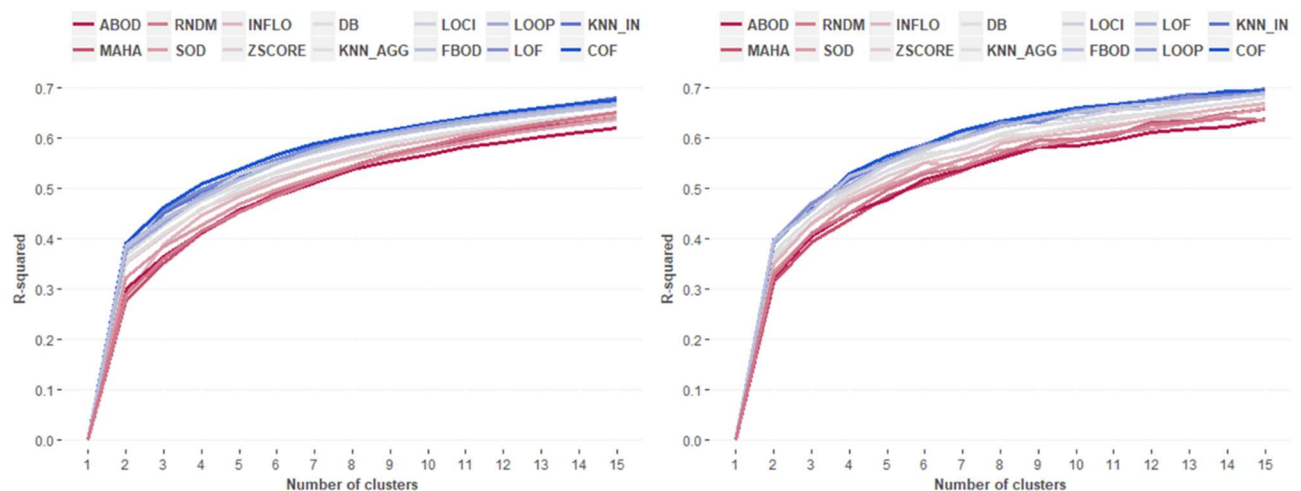


Figure 29: R^2 for *BreastCancerDiagnostic* with hierarchical clustering (left) and k-means (right) with 15% removed observations

9.3.4. GesturePhase

For hierarchical clustering the greatest improvement of clustering results is achieved with LOCI treated dataset (15%), returning an R^2 of 0.279, 0.05 points better than RNDM treated dataset.

The seven best solutions are predominately achieved with the density-based approaches, COF, FBOD and LOCI, the distance-based approaches KNN_AGG, KNN_IN and DB and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.007 in average, when removing additional 5% observations.

The seven worst solutions are returned by the high-dimensional approach ABOD, the statistical approaches ZSCORE and MAHA and the density-based approach INFLO, LOF and LOOP. For these solutions, R^2 increase with 0.003 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 16887 (15%) followed by DB and ZSCORE with an SSW of 17583 and 17915, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOCI	LOCI	LOCI	0.252	0.272	0.279	34161	27679	23766
2	KNN_AGG	DB	KNN_IN	0.244	0.253	0.271	30731	25067	20870
3	SOD	SOD	SOD	0.240	0.249	0.253	35354	27143	20345
4	COF	KNN_AGG	LOOP	0.239	0.248	0.249	49430	22462	36909
5	RNDM	KNN_IN	DB	0.238	0.246	0.249	48081	26498	17583
6	DB	ABOD	KNN_AGG	0.238	0.238	0.248	39290	31389	16887
7	KNN_IN	RNDM	LOF	0.236	0.237	0.246	34999	44091	40255
8	ZSCORE	FBOD	FBOD	0.235	0.236	0.240	31911	38688	38520
9	LOOP	LOOP	ABOD	0.233	0.236	0.240	45500	42546	29900
10	FBOD	LOF	COF	0.231	0.236	0.238	47038	44151	44883
11	ABOD	MAHA	MAHA	0.230	0.235	0.237	34892	38873	34122
12	MAHA	COF	ZSCORE	0.225	0.235	0.231	44834	47775	17915
13	LOF	ZSCORE	RNDM	0.222	0.235	0.229	48442	23599	42140
14	INFLO	INFLO	INFLO	0.219	0.219	0.225	34523	28038	24406

Table 28: Hierarchical clustering results for GesturePhase

For k-means clustering the greatest improvement of clustering results is achieved with LOCI treated dataset (15%), returning an R^2 of 0.322, 0.052 points better than RNDM treated dataset.

The seven best solutions are predominately returned by the density-based approaches LOCI and FBOD, the distance-based approaches KNN_AGG and DB, the statistical test ZSCORE and the high-dimensional approach SOD. For these solutions, R^2 increase with 0.010 in average when removing additional 5% observations.

The seven worst solutions are predominately returned by the density-based approach LOOP, LOF, COF and INFLO and the statistical approach MAHA. Of these solutions, COF returns worse results than with RNDM treated dataset. For these solutions, R^2 increase with 0.003 points in average, when removing additional 5% observations.

Inspecting SSW, the greatest variance reduction is achieved by KNN_AGG with an SSW of 15712 (15%) followed by DB and ZSCORE with an SSW of 16411 and 16412, respectively.

Rank	Method			R^2			SSW		
	5%	10%	15%	5%	10%	15%	5%	10%	15%
1	LOCI	LOCI	LOCI	0.295	0.308	0.322	32179	26301	22366
2	KNN_AGG	KNN_IN	KNN_IN	0.283	0.296	0.313	29160	24760	19688
3	KNN_IN	ZSCORE	KNN_AGG	0.282	0.291	0.300	32873	21849	15712
4	ZSCORE	DB	DB	0.281	0.290	0.299	30006	23829	16411
5	SOD	KNN_AGG	ZSCORE	0.279	0.289	0.295	33557	21242	16415
6	DB	SOD	SOD	0.274	0.286	0.294	37404	25786	19212
7	ABOD	FBOD	FBOD	0.273	0.279	0.292	31315	38442	34009
8	MAHA	LOOP	LOOP	0.272	0.278	0.283	42086	40168	35241
9	LOOP	MAHA	MAHA	0.272	0.277	0.281	43175	36760	32159
10	INFLO	ABOD	LOF	0.272	0.276	0.280	32190	23903	38404
11	LOF	LOF	ABOD	0.271	0.273	0.280	45431	41968	20118
12	FBOD	INFLO	COF	0.270	0.271	0.273	43761	26141	42842
13	RNDM	COF	INFLO	0.269	0.270	0.273	46479	45618	22908
14	COF	RNDM	RNDM	0.266	0.269	0.270	47723	43157	41539

Table 29: K-means clustering results for GesturePhase

For *GesturePhase* it is concluded that k-means clustering returns the best results, with an R^2 of 0.322 compared to an R^2 of 0.279 for hierarchical clustering.

In general, the clustering results are poor, with the lowest R^2 for all clustering results across datasets. With 5% removed observations, the best outcome is only 0.014 points better than RNDM treated dataset for hierarchical clustering. Moreover, removing 5% additional observations only improves the clustering results with between 0.7% to 1% for the two clustering algorithms, indicating that *GesturePhase* has dense characteristics.

When asked how outlier detection can improve clustering results, the greatest impact is also observed for k-means clustering, improving the overall clustering results with 0.052 points compared to 0.050 for hierarchical clustering. Interestingly, LOCI has the best outcome for both clustering algorithms, leaving behind the other density-based approaches that LOCI shares strong similarities with.

Inspecting SSW, it is concluded that KNN_AGG, ZSCORE and DB all excel in creating dense clusters, with the lowest SSW observed for both clustering methods. However, following R^2 , better results are achieved with LOCI. Thus, KNN_AGG, ZSCORE and DB creates the densest clusters in absolute figures but fails to explain the proportion of remaining variance.

Despite of *GesturePhase* being a high-dimensional dataset, ABOD returns mixed results but still improves relatively when compared to results for *PimaIndians*, *Frogs* and *HeartDisease*. Again, SOD deviates strongly from the results of ABOD, reflection their underlying differences for judging outlierness.

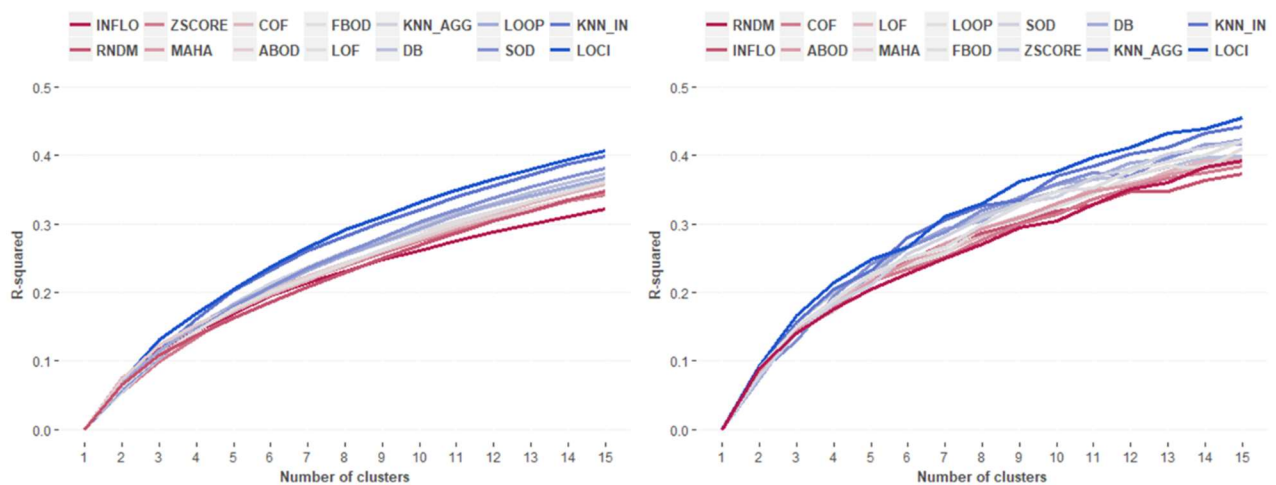


Figure 30: R^2 for *GesturePhase* with hierarchical clustering (left) and k-means (right) with 15% removed observations

9.4. FIGURES FOR INCOHERENT RESULTS

9.4.1. BreastCancerDiagnostic

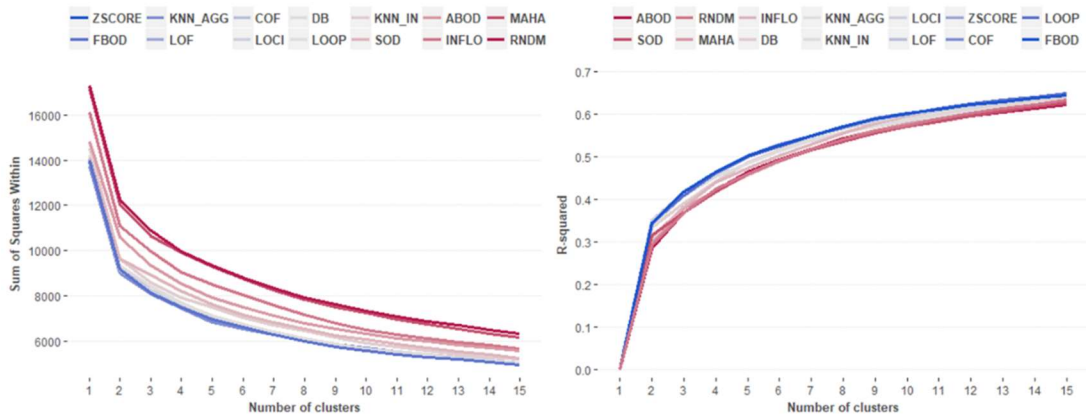


Figure 31: Hierarchical clustering for BreastCancerDiagnostic with 5% of observations removed

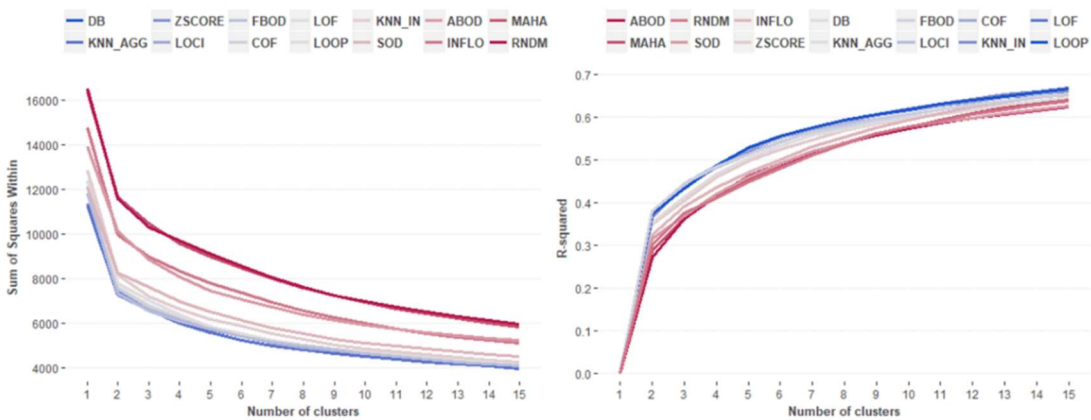


Figure 32: Hierarchical clustering for BreastCancerDiagnostic with 10% of observations removed

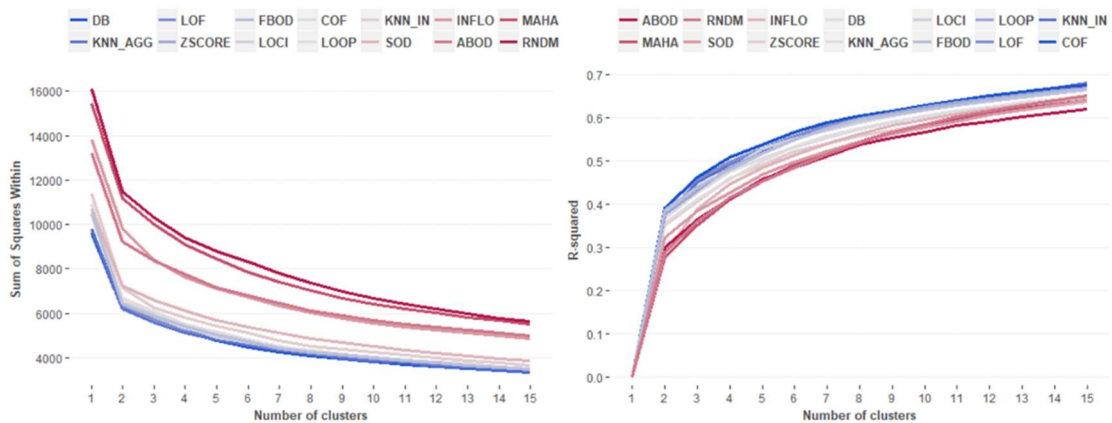


Figure 33: Hierarchical clustering for BreastCancerDiagnostic with 15% of observations removed

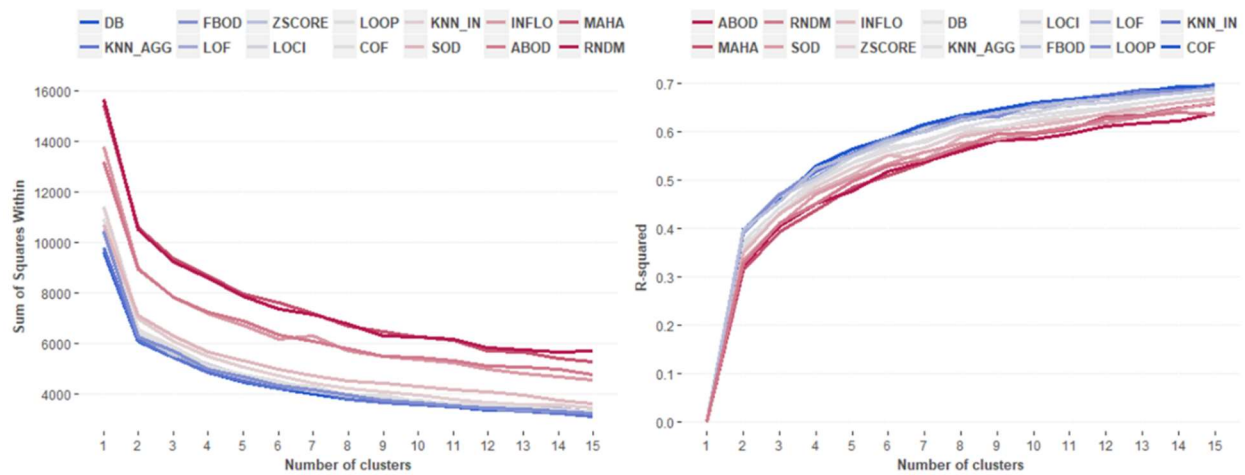


Figure 34: K-means clustering for BreastCancerDiagnostic with 5% of observations removed

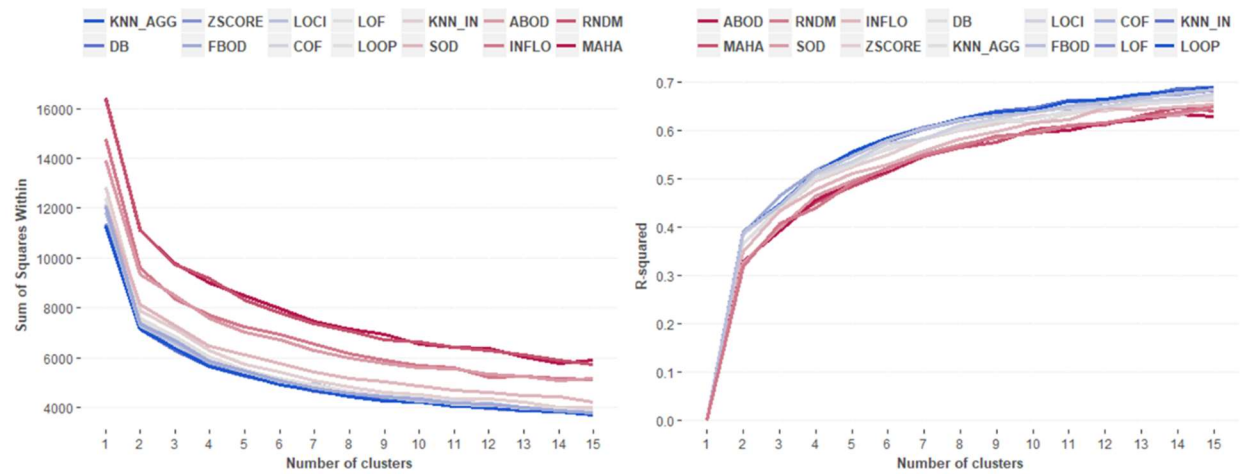


Figure 35: K-means clustering for BreastCancerDiagnostic with 10% of observations removed

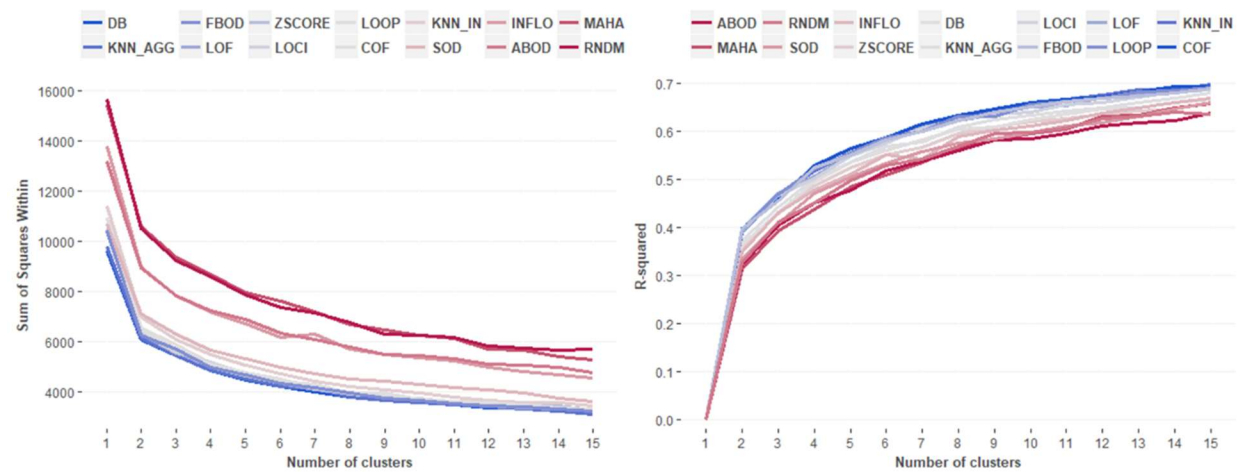


Figure 36: K-means clustering for BreastCancerDiagnostic with 15% of observations removed

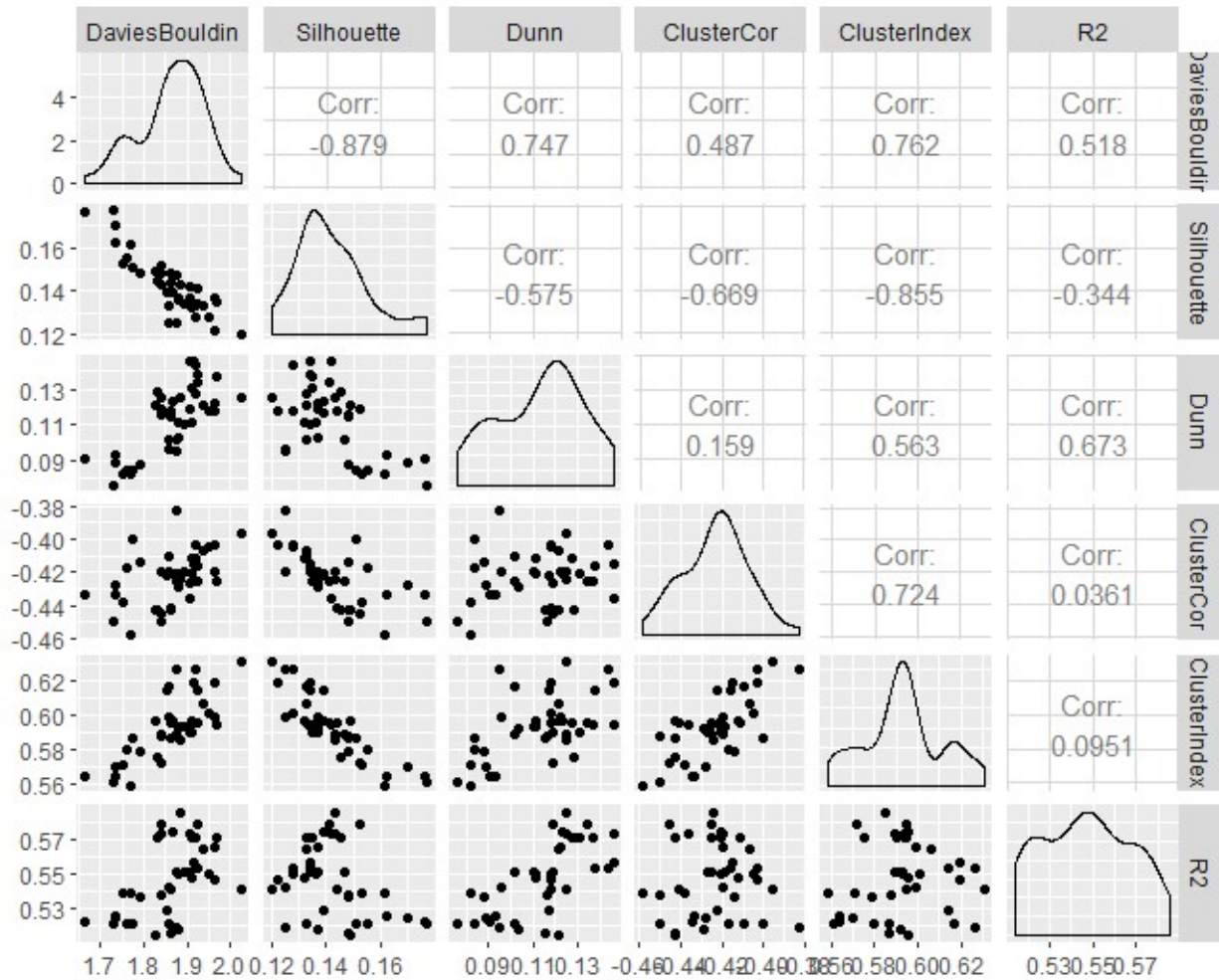


Figure 37: Correlation matrix plot for clustering evaluation metrics for BreastCancerDiagnostic with hierarchical clustering

Comment on interpretation: Figure 37 shows the correlation between the results for six different cluster evaluation metrics averaged out over cluster solutions for 2:15 clusters for k-means and hierarchical clustering with 5%, 10% and 15% outlier removal.

As an example, it can be seen in the upper triangular part of the matrix that the results for cluster index and Silhouette coefficient have a correlation of -0.85, which means that increasing Silhouette coefficient correlates negatively with the cluster index. Thus, a greater Silhouette coefficient indicating better clustering results equals a lower cluster index, also indicating better clustering results. In the lower triangular part of the matrix the observations are plotted for Silhouette coefficient and cluster index, displaying the negative correlation between the two evaluation metrics. On the y-axis it can be seen that a Silhouette coefficient of ≈ 0.63 is equal to a cluster index of ≈ 0.12 (corresponding x-axis). However, there is little to no correlation (0.095) between R^2 and cluster index, meaning that solution lacks coherence.

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
FBOD	1.906	0.137	0.119	-0.427	0.590	0.551	6284
LOOP	1.877	0.147	0.102	-0.425	0.589	0.551	6403
COF	1.891	0.134	0.110	-0.419	0.593	0.551	6345
ZSCORE	1.949	0.128	0.118	-0.405	0.601	0.550	6284
LOF	1.879	0.136	0.111	-0.421	0.593	0.550	6339
LOCI	1.910	0.132	0.111	-0.412	0.596	0.548	6379
KNN_IN	1.856	0.125	0.096	-0.420	0.599	0.542	6654
KNN_AGG	1.860	0.144	0.118	-0.441	0.595	0.541	6302
DB	1.840	0.148	0.116	-0.450	0.588	0.538	6381
INFLO	1.789	0.148	0.088	-0.414	0.579	0.537	7470
MAHA	1.732	0.170	0.089	-0.428	0.570	0.524	8167
RNDM	1.775	0.151	0.084	-0.400	0.587	0.521	8288
SOD	1.876	0.125	0.095	-0.383	0.626	0.519	6790
ABOD	1.881	0.137	0.103	-0.429	0.592	0.518	7143

Table 30: Hierarchical clustering results for BreastCancerDiagnostic with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOOP	1.840	0.143	0.125	-0.420	0.589	0.573	5305
LOF	1.908	0.135	0.131	-0.421	0.590	0.571	5231
KNN_IN	1.829	0.145	0.128	-0.443	0.575	0.571	5515
COF	1.920	0.133	0.127	-0.411	0.595	0.571	5208
LOCI	1.961	0.137	0.122	-0.420	0.599	0.565	5156
FBOD	1.937	0.133	0.121	-0.407	0.606	0.564	5158
DB	1.924	0.134	0.138	-0.416	0.614	0.553	5045
KNN_AGG	1.924	0.134	0.138	-0.416	0.614	0.553	5045
ZSCORE	1.963	0.122	0.118	-0.403	0.619	0.547	5143
INFLO	1.750	0.153	0.082	-0.438	0.571	0.539	6804
SOD	1.856	0.133	0.102	-0.410	0.617	0.521	5788
RNDM	1.729	0.177	0.076	-0.450	0.561	0.521	7922
MAHA	1.760	0.155	0.084	-0.417	0.580	0.521	7868
ABOD	1.860	0.148	0.115	-0.443	0.587	0.516	6733

Table 31: Hierarchical clustering results for BreastCancerDiagnostic with 10% removed observations

	Avg DaviesBouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.885	0.143	0.125	-0.424	0.585	0.586	4520
KNN_IN	1.839	0.152	0.119	-0.445	0.572	0.579	4793
LOF	1.925	0.141	0.134	-0.425	0.595	0.579	4411
LOOP	1.864	0.139	0.123	-0.421	0.596	0.574	4556
FBOD	1.906	0.142	0.146	-0.436	0.594	0.573	4469
LOCI	1.967	0.135	0.137	-0.425	0.594	0.571	4479
KNN_AGG	1.913	0.134	0.146	-0.415	0.619	0.557	4332
DB	1.918	0.128	0.144	-0.403	0.626	0.553	4289
ZSCORE	2.026	0.120	0.125	-0.396	0.631	0.541	4454
INFLO	1.769	0.161	0.082	-0.458	0.559	0.539	6362
SOD	1.853	0.139	0.117	-0.422	0.614	0.529	5045
RNDM	1.735	0.162	0.093	-0.433	0.564	0.525	7653
MAHA	1.665	0.176	0.091	-0.434	0.564	0.522	7383
ABOD	1.828	0.149	0.121	-0.443	0.596	0.514	6410

Table 32: Hierarchical clustering results for BreastCancerDiagnostic with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOOP	1.765	0.160	0.106	-0.416	0.571	0.598	4989
KNN_IN	1.753	0.162	0.106	-0.419	0.566	0.597	5177
LOF	1.789	0.159	0.107	-0.415	0.576	0.596	4929
COF	1.760	0.168	0.112	-0.426	0.575	0.596	4905
FBOD	1.794	0.158	0.110	-0.411	0.581	0.593	4898
LOCI	1.804	0.157	0.119	-0.421	0.583	0.586	4901
KNN_AGG	1.792	0.157	0.124	-0.430	0.591	0.579	4751
DB	1.788	0.164	0.130	-0.441	0.588	0.578	4760
ZSCORE	1.806	0.153	0.108	-0.415	0.597	0.575	4831
INFLO	1.744	0.172	0.071	-0.433	0.557	0.562	6469
SOD	1.816	0.152	0.090	-0.402	0.602	0.546	5485
RNDM	1.807	0.163	0.069	-0.400	0.557	0.544	7463
MAHA	1.748	0.165	0.066	-0.390	0.566	0.544	7494
ABOD	1.853	0.146	0.088	-0.407	0.588	0.542	6372

Table 33: K-means clustering results for BreastCancerDiagnostic with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOOP	1.768	0.161	0.107	-0.418	0.571	0.597	4999
COF	1.760	0.169	0.113	-0.427	0.572	0.597	4893
KNN_IN	1.692	0.182	0.110	-0.440	0.561	0.597	5186
LOF	1.773	0.167	0.107	-0.423	0.573	0.596	4928
FBOD	1.807	0.157	0.108	-0.415	0.581	0.593	4924
LOCI	1.816	0.153	0.118	-0.415	0.586	0.586	4906
DB	1.808	0.158	0.130	-0.430	0.592	0.578	4764
KNN_AGG	1.815	0.155	0.126	-0.420	0.597	0.575	4798
ZSCORE	1.788	0.156	0.118	-0.416	0.598	0.572	4866
INFLO	1.788	0.168	0.070	-0.421	0.557	0.565	6432
SOD	1.831	0.161	0.094	-0.418	0.595	0.546	5494
MAHA	1.799	0.159	0.065	-0.381	0.569	0.544	7479
ABOD	1.859	0.154	0.079	-0.420	0.582	0.544	6351
RNDM	1.794	0.162	0.071	-0.393	0.567	0.539	7572

Table 34: K-means clustering results for BreastCancerDiagnostic with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.749	0.164	0.115	-0.420	0.571	0.607	4288
KNN_IN	1.705	0.174	0.112	-0.438	0.559	0.605	4503
LOOP	1.749	0.172	0.117	-0.439	0.570	0.603	4254
LOF	1.735	0.176	0.128	-0.447	0.573	0.600	4183
FBOD	1.810	0.161	0.117	-0.428	0.578	0.600	4176
LOCI	1.757	0.168	0.119	-0.431	0.579	0.597	4213
KNN_AGG	1.755	0.164	0.140	-0.431	0.595	0.584	4069
DB	1.786	0.164	0.142	-0.440	0.597	0.577	4068
ZSCORE	1.807	0.157	0.132	-0.427	0.600	0.568	4191
INFLO	1.765	0.168	0.072	-0.422	0.556	0.565	6007
SOD	1.797	0.161	0.104	-0.414	0.599	0.553	4782
RNDM	1.760	0.175	0.073	-0.419	0.556	0.548	7086
MAHA	1.773	0.162	0.068	-0.387	0.563	0.543	7071
ABOD	1.841	0.151	0.092	-0.412	0.591	0.538	6101

Table 35: K-means clustering results for BreastCancerDiagnostic with 15% removed observations

9.4.2. BreastCancerOriginal

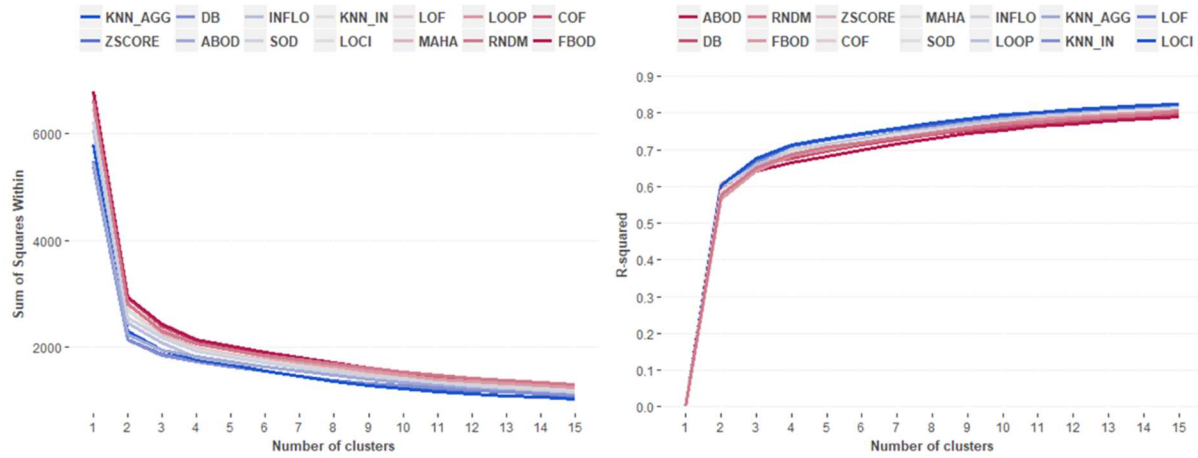


Figure 38: Hierarchical clustering for BreastCancerOriginal with 5% of observations removed

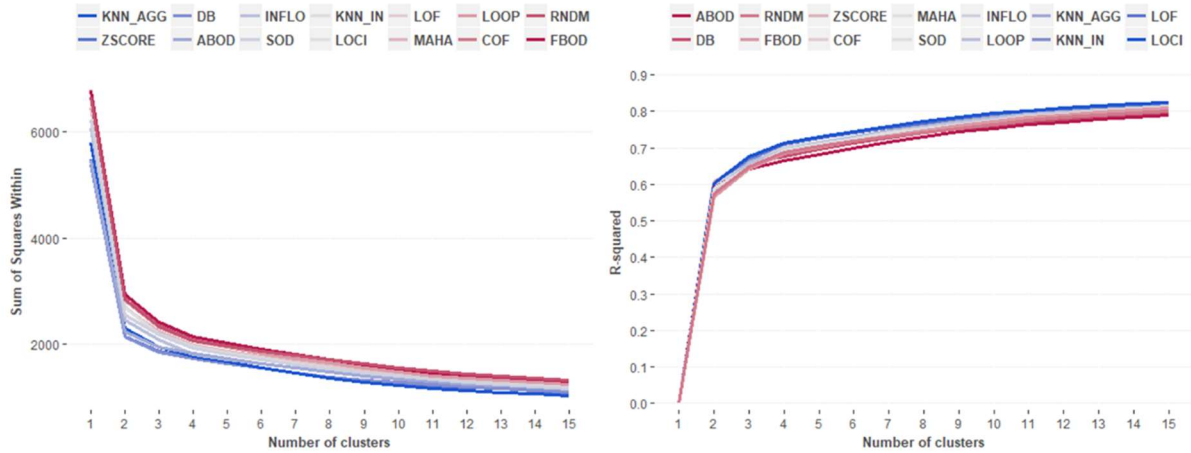


Figure 39: Hierarchical clustering for BreastCancerOriginal with 10% of observations removed

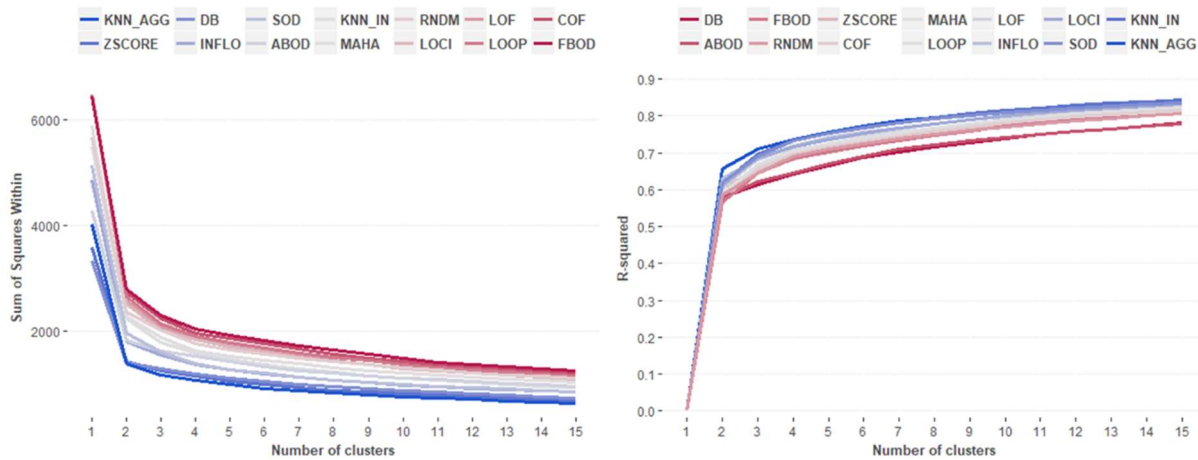


Figure 40: Hierarchical clustering for BreastCancerOriginal with 15% of observations removed

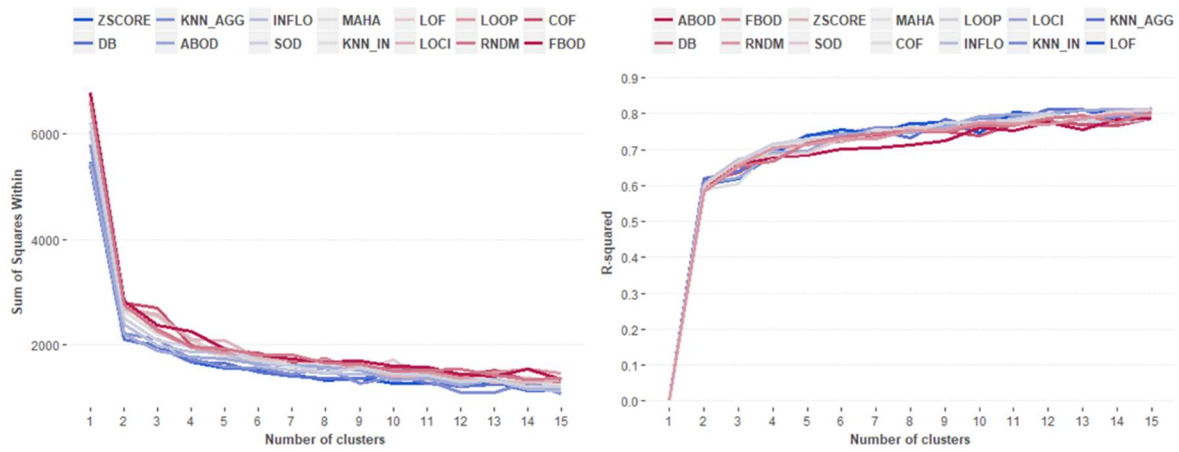


Figure 41: K-means clustering for BreastCancerOriginal with 5% of observations removed

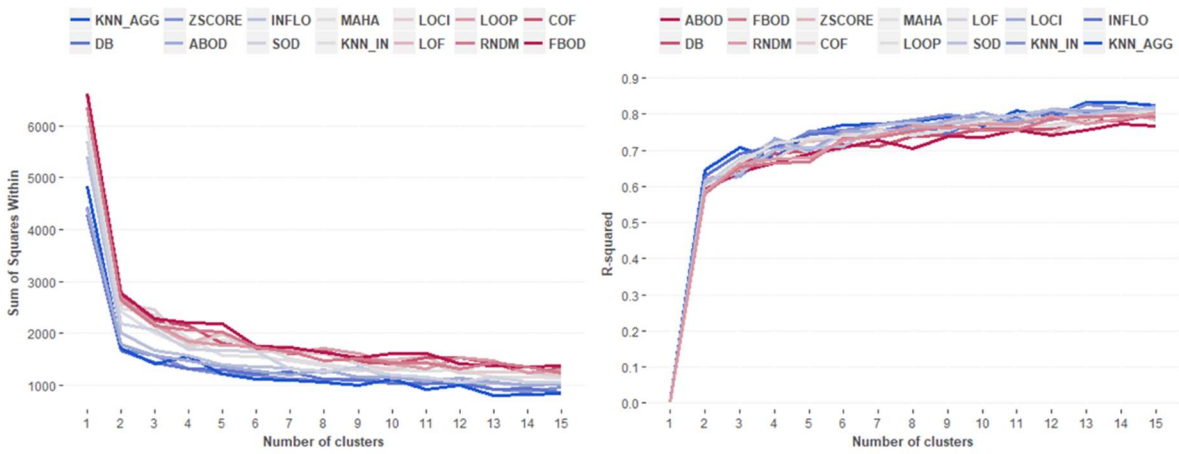


Figure 42: K-means clustering for BreastCancerOriginal with 10% of observations removed

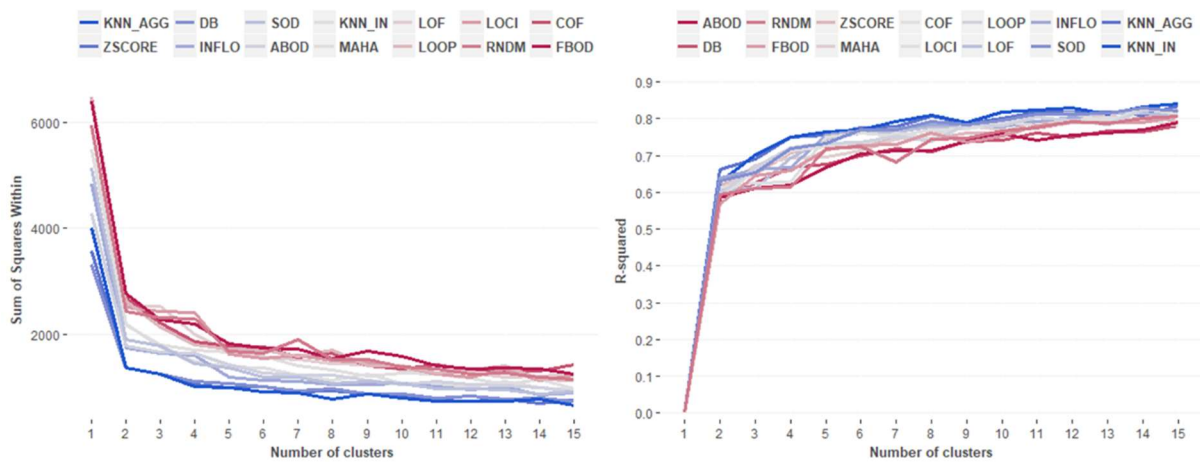


Figure 43: K-means clustering for BreastCancerOriginal with 15% of observations removed

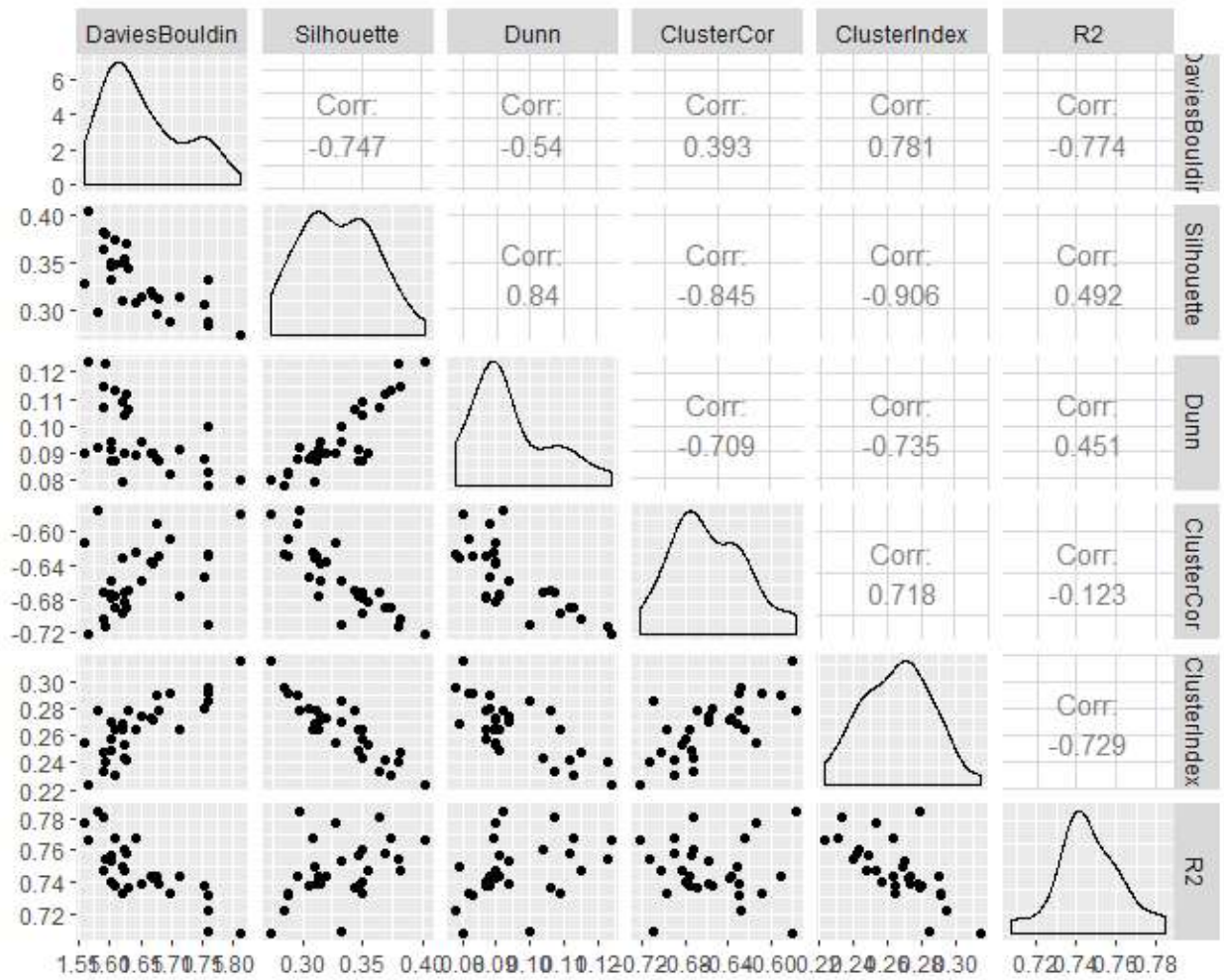


Figure 44: Correlation matrix plot for clustering evaluation metrics for BreastCancerOriginal with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	1.623	0.350	0.104	-0.672	0.244	0.761	1620
LOF	1.626	0.369	0.112	-0.691	0.242	0.758	1640
KNN_IN	1.601	0.346	0.091	-0.674	0.249	0.757	1607
KNN_AGG	1.601	0.332	0.094	-0.659	0.270	0.753	1433
LOOP	1.624	0.354	0.090	-0.684	0.253	0.747	1714
INFLO	1.713	0.313	0.091	-0.676	0.264	0.744	1552
SOD	1.667	0.319	0.090	-0.637	0.273	0.743	1602
MAHA	1.669	0.315	0.090	-0.638	0.272	0.742	1665
COF	1.602	0.350	0.087	-0.679	0.257	0.740	1762
RNDM	1.680	0.311	0.087	-0.630	0.278	0.739	1754
ZSCORE	1.753	0.306	0.088	-0.655	0.280	0.737	1441
FBOD	1.619	0.349	0.109	-0.697	0.265	0.733	1809
DB	1.759	0.287	0.083	-0.630	0.291	0.732	1443
ABOD	1.759	0.284	0.078	-0.627	0.295	0.722	1518

Table 36: Hierarchical clustering results for BreastCancerOriginal with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	1.623	0.350	0.104	-0.672	0.244	0.761	1620
LOF	1.626	0.369	0.112	-0.691	0.242	0.758	1640
KNN_IN	1.601	0.346	0.091	-0.674	0.249	0.757	1607
KNN_AGG	1.601	0.332	0.094	-0.659	0.270	0.753	1433
LOOP	1.624	0.354	0.090	-0.684	0.253	0.747	1714
INFLO	1.713	0.313	0.091	-0.676	0.264	0.744	1552
SOD	1.667	0.319	0.090	-0.637	0.273	0.743	1602
MAHA	1.669	0.315	0.090	-0.638	0.272	0.742	1665
COF	1.602	0.350	0.087	-0.679	0.257	0.740	1762
ZSCORE	1.753	0.306	0.088	-0.655	0.280	0.737	1441
FBOD	1.608	0.347	0.087	-0.676	0.265	0.737	1781
RNDM	1.699	0.287	0.082	-0.608	0.291	0.733	1772
DB	1.759	0.287	0.083	-0.630	0.291	0.732	1443
ABOD	1.759	0.284	0.078	-0.627	0.295	0.722	1518

Table 37: Hierarchical clustering results for BreastCancerOriginal with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.580	0.298	0.092	-0.576	0.279	0.785	863
KNN_IN	1.588	0.364	0.107	-0.673	0.233	0.781	1289
SOD	1.560	0.327	0.090	-0.613	0.254	0.777	1145
LOCI	1.608	0.373	0.113	-0.691	0.231	0.768	1499
INFLO	1.643	0.308	0.089	-0.624	0.264	0.768	1123
LOF	1.565	0.403	0.124	-0.722	0.224	0.766	1519
LOOP	1.593	0.380	0.123	-0.713	0.240	0.755	1577
MAHA	1.619	0.310	0.079	-0.631	0.269	0.749	1375
COF	1.590	0.382	0.115	-0.703	0.248	0.747	1634
ZSCORE	1.676	0.296	0.088	-0.590	0.290	0.743	920
RNDM	1.651	0.314	0.094	-0.658	0.274	0.739	1479
FBOD	1.631	0.343	0.106	-0.669	0.279	0.736	1698
ABOD	1.759	0.332	0.100	-0.710	0.285	0.709	1243
DB	1.813	0.273	0.080	-0.579	0.315	0.707	970

Table 38: Hierarchical clustering results for BreastCancerOriginal with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOF	1.638	0.259	0.054	-0.481	0.315	0.748	1705
KNN_AGG	1.722	0.254	0.070	-0.474	0.335	0.746	1468
KNN_IN	1.703	0.238	0.056	-0.436	0.339	0.746	1682
LOCI	1.658	0.241	0.054	-0.447	0.331	0.746	1722
INFLO	1.773	0.253	0.066	-0.483	0.320	0.745	1545
LOOP	1.657	0.282	0.070	-0.495	0.320	0.744	1732
COF	1.691	0.290	0.066	-0.536	0.306	0.741	1760
MAHA	1.795	0.271	0.073	-0.510	0.308	0.740	1672
SOD	1.784	0.254	0.061	-0.483	0.312	0.740	1617
ZSCORE	1.814	0.232	0.066	-0.482	0.323	0.738	1435
RNDM	1.774	0.314	0.092	-0.566	0.296	0.737	1738
FBOD	1.775	0.255	0.061	-0.509	0.316	0.736	1791
DB	1.809	0.239	0.063	-0.479	0.336	0.729	1458
ABOD	1.844	0.253	0.064	-0.492	0.331	0.720	1529

Table 39: K-means clustering results for BreastCancerOriginal with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.627	0.296	0.087	-0.517	0.306	0.768	1121
INFLO	1.718	0.257	0.071	-0.487	0.307	0.757	1312
KNN_IN	1.636	0.259	0.059	-0.463	0.322	0.757	1545
LOCI	1.589	0.280	0.061	-0.476	0.320	0.754	1627
SOD	1.779	0.245	0.056	-0.456	0.307	0.751	1418
LOF	1.640	0.292	0.070	-0.490	0.319	0.751	1652
LOOP	1.667	0.289	0.071	-0.516	0.322	0.746	1675
MAHA	1.740	0.288	0.069	-0.541	0.291	0.745	1528
COF	1.665	0.282	0.062	-0.514	0.312	0.743	1703
ZSCORE	1.745	0.240	0.067	-0.444	0.337	0.737	1168
RNDM	1.778	0.267	0.066	-0.517	0.311	0.736	1678
FBOD	1.770	0.274	0.067	-0.515	0.333	0.731	1774
DB	1.833	0.258	0.079	-0.491	0.324	0.730	1160
ABOD	1.857	0.255	0.066	-0.495	0.324	0.714	1255

Table 40: K-means clustering results for BreastCancerOriginal with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	1.625	0.324	0.090	-0.555	0.267	0.783	1279
KNN_AGG	1.668	0.252	0.067	-0.446	0.312	0.779	888
SOD	1.704	0.255	0.055	-0.445	0.290	0.767	1195
INFLO	1.718	0.247	0.064	-0.454	0.325	0.760	1160
LOF	1.611	0.268	0.055	-0.473	0.308	0.755	1586
LOOP	1.619	0.298	0.073	-0.544	0.303	0.753	1593
LOCI	1.559	0.284	0.064	-0.466	0.337	0.751	1608
COF	1.624	0.320	0.081	-0.575	0.300	0.747	1640
MAHA	1.780	0.265	0.066	-0.486	0.310	0.744	1401
ZSCORE	1.718	0.256	0.072	-0.478	0.320	0.744	918
FBOD	1.730	0.282	0.071	-0.547	0.317	0.732	1711
RNDM	1.822	0.229	0.062	-0.438	0.355	0.726	1631
DB	1.764	0.249	0.077	-0.469	0.335	0.714	949
ABOD	1.788	0.253	0.069	-0.498	0.348	0.709	1246

Table 41: K-means clustering results for BreastCancerOriginal with 15% removed observations

9.4.3. ForestFires

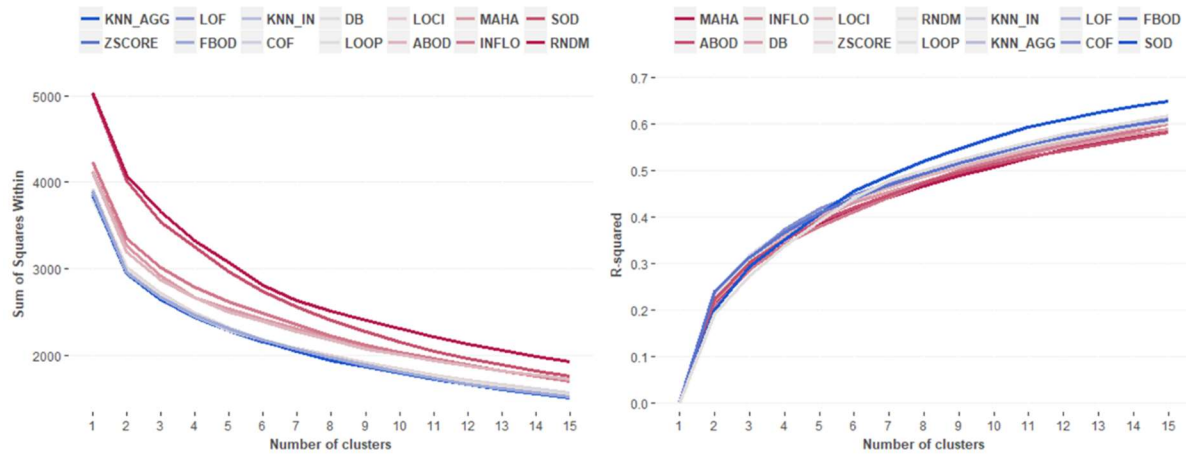


Figure 45: Hierarchical clustering for ForestFires with 5% of observations removed

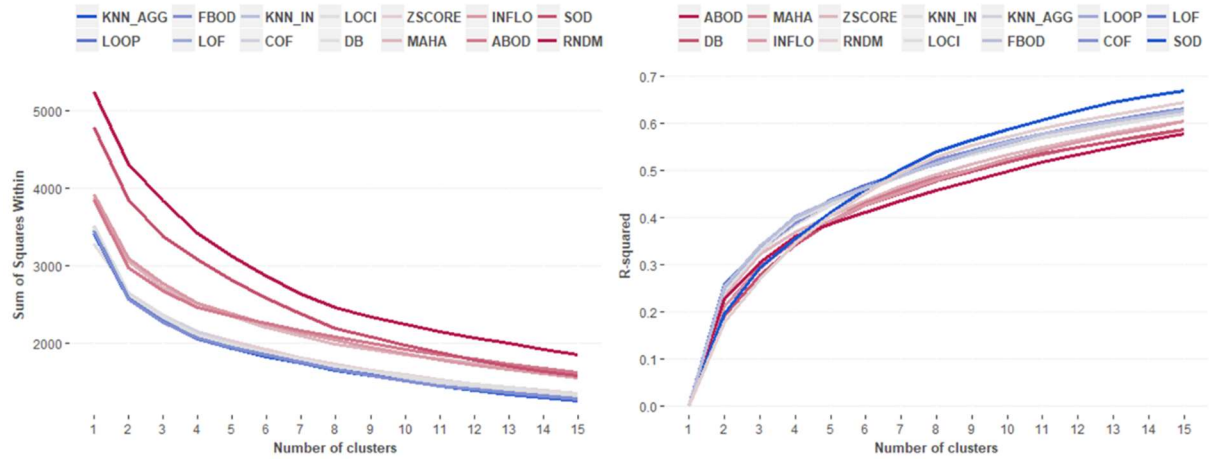


Figure 46: Hierarchical clustering for ForestFires with 10% of observations removed

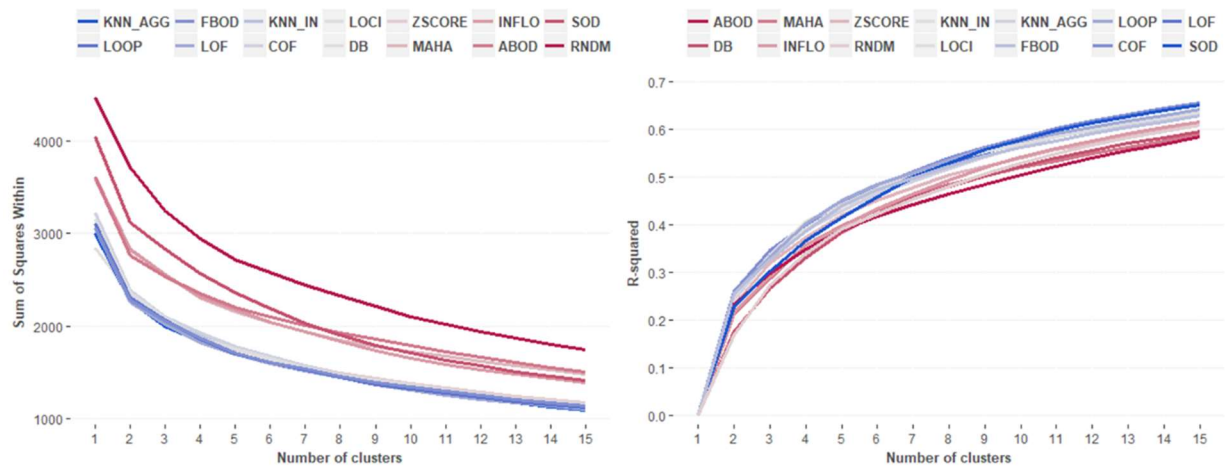


Figure 47: Hierarchical clustering for ForestFires with 15% of observations removed

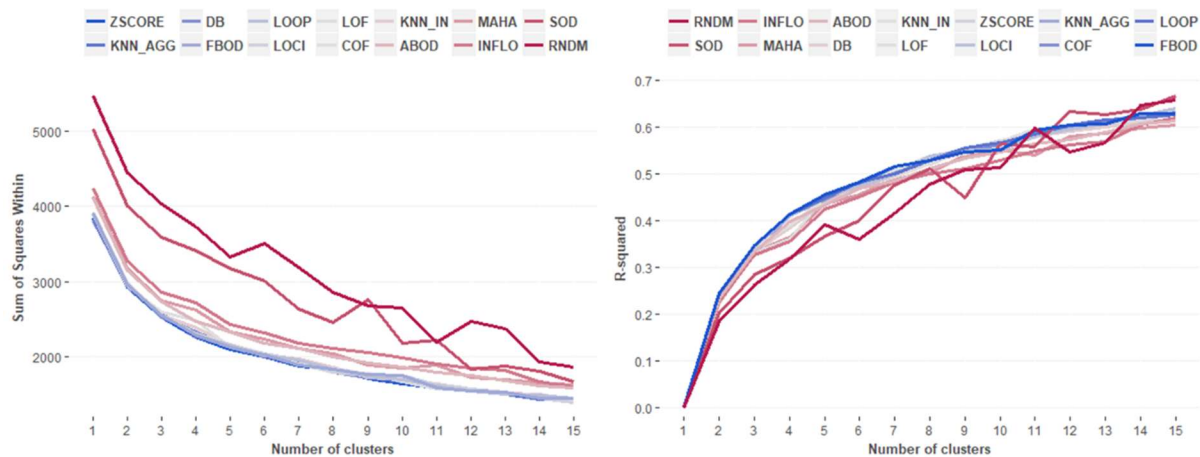


Figure 48: K-means clustering for ForestFires with 5% of observations removed

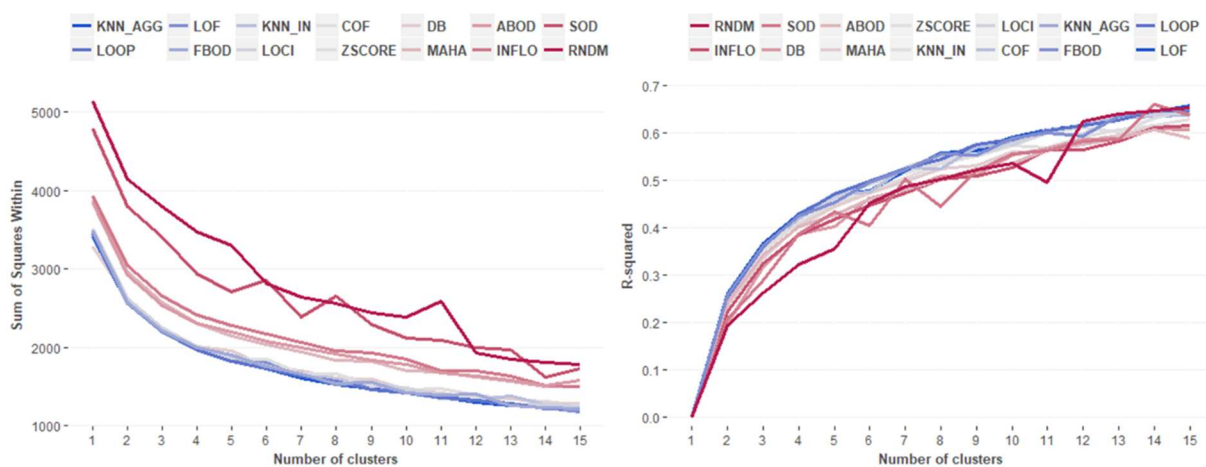


Figure 49: K-means clustering for ForestFires with 10% of observations removed

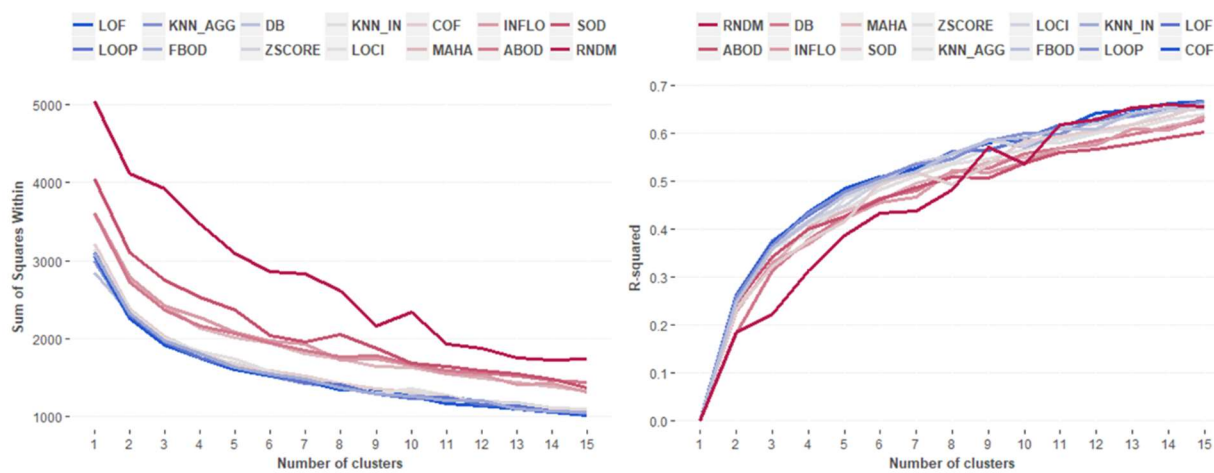


Figure 50: K-means clustering for ForestFires with 15% of observations removed

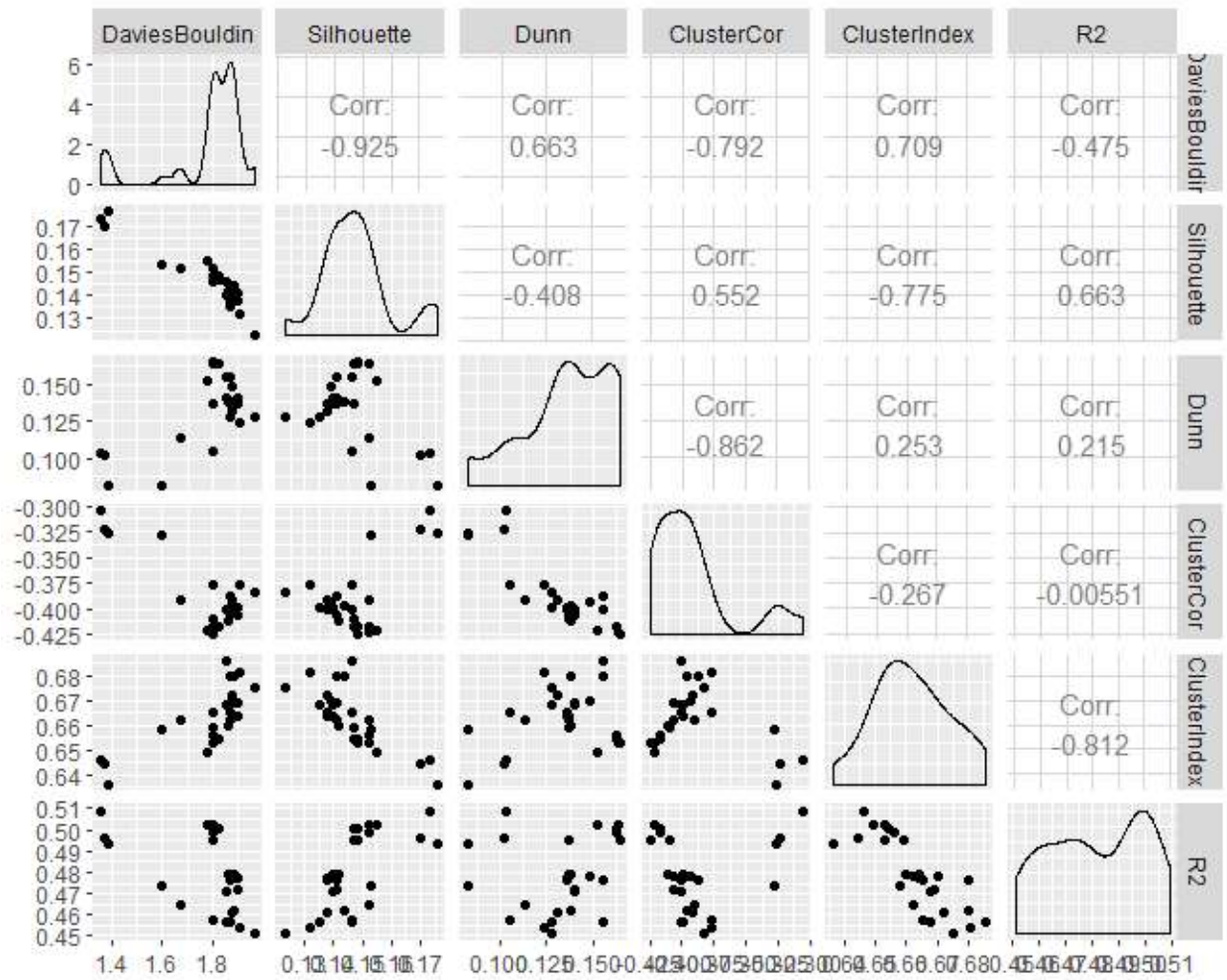


Figure 51: Correlation matrix plot for clustering evaluation metrics for ForestFires with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
SOD	1.375	0.170	0.102	-0.323	0.644	0.496	2528
FBOD	1.863	0.142	0.138	-0.411	0.660	0.479	2029
COF	1.879	0.140	0.137	-0.399	0.664	0.479	2047
LOF	1.866	0.141	0.137	-0.407	0.662	0.478	2029
KNN_AGG	1.871	0.139	0.148	-0.394	0.670	0.478	2013
KNN_IN	1.895	0.138	0.136	-0.399	0.664	0.477	2044
LOOP	1.869	0.138	0.136	-0.400	0.665	0.476	2050
RNDM	1.601	0.153	0.082	-0.328	0.658	0.474	2650
ZSCORE	1.897	0.141	0.140	-0.407	0.669	0.472	2024
LOCI	1.851	0.140	0.140	-0.400	0.668	0.471	2059
DB	1.884	0.144	0.138	-0.397	0.680	0.462	2048
INFLO	1.796	0.146	0.105	-0.377	0.665	0.458	2293
ABOD	1.870	0.135	0.128	-0.399	0.668	0.457	2235
MAHA	1.906	0.132	0.124	-0.377	0.681	0.454	2255

Table 42: Hierarchical clustering results for ForestFires with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
SOD	1.361	0.173	0.103	-0.305	0.646	0.509	2353
LOF	1.803	0.152	0.164	-0.422	0.653	0.502	1724
COF	1.780	0.155	0.152	-0.421	0.649	0.502	1753
LOOP	1.820	0.147	0.163	-0.417	0.654	0.501	1719
FBOD	1.825	0.148	0.163	-0.417	0.654	0.501	1719
KNN_AGG	1.800	0.152	0.163	-0.417	0.656	0.499	1710
LOCI	1.803	0.147	0.137	-0.410	0.659	0.495	1764
KNN_IN	1.802	0.148	0.165	-0.425	0.653	0.495	1752
RNDM	1.390	0.176	0.082	-0.327	0.636	0.493	2661
ZSCORE	1.869	0.141	0.155	-0.387	0.680	0.476	1789
INFLO	1.669	0.152	0.113	-0.391	0.662	0.465	2096
MAHA	1.876	0.138	0.131	-0.392	0.672	0.461	2088
DB	1.852	0.146	0.155	-0.400	0.686	0.457	1781
ABOD	1.965	0.123	0.128	-0.383	0.675	0.451	2116

Table 43: Hierarchical clustering results for ForestFires with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
SOD	1.361	0.173	0.103	-0.305	0.646	0.509	2353
LOF	1.803	0.152	0.164	-0.422	0.653	0.502	1724
COF	1.780	0.155	0.152	-0.421	0.649	0.502	1753
LOOP	1.820	0.147	0.163	-0.417	0.654	0.501	1719
FBOD	1.825	0.148	0.163	-0.417	0.654	0.501	1719
KNN_AGG	1.800	0.152	0.163	-0.417	0.656	0.499	1710
LOCI	1.803	0.147	0.137	-0.410	0.659	0.495	1764
KNN_IN	1.802	0.148	0.165	-0.425	0.653	0.495	1752
RNDM	1.390	0.176	0.082	-0.327	0.636	0.493	2661
ZSCORE	1.869	0.141	0.155	-0.387	0.680	0.476	1789
INFLO	1.669	0.152	0.113	-0.391	0.662	0.465	2096
MAHA	1.876	0.138	0.131	-0.392	0.672	0.461	2088
DB	1.852	0.146	0.155	-0.400	0.686	0.457	1781
ABOD	1.965	0.123	0.128	-0.383	0.675	0.451	2116

Table 44: Hierarchical clustering results for ForestFires with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
FBOD	1.705	0.168	0.104	-0.420	0.640	0.512	1924
LOF	1.729	0.171	0.111	-0.424	0.642	0.510	1905
LOOP	1.697	0.172	0.100	-0.419	0.645	0.509	1919
KNN_AGG	1.728	0.167	0.111	-0.424	0.645	0.509	1892
COF	1.715	0.168	0.103	-0.420	0.642	0.509	1926
LOCI	1.734	0.166	0.107	-0.422	0.644	0.508	1918
ZSCORE	1.738	0.166	0.107	-0.421	0.651	0.504	1903
KNN_IN	1.727	0.167	0.099	-0.407	0.653	0.504	1939
DB	1.752	0.167	0.108	-0.420	0.652	0.499	1907
MAHA	1.786	0.160	0.095	-0.398	0.654	0.491	2099
ABOD	1.800	0.158	0.091	-0.405	0.652	0.488	2108
INFLO	1.753	0.159	0.062	-0.377	0.648	0.484	2183
SOD	1.798	0.163	0.039	-0.273	0.650	0.480	2608
RNDM	1.802	0.159	0.037	-0.281	0.635	0.466	2917

Table 45: K-means clustering results for ForestFires with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOF	1.687	0.175	0.112	-0.431	0.639	0.527	1638
FBOD	1.686	0.174	0.122	-0.432	0.640	0.524	1645
KNN_AGG	1.687	0.171	0.122	-0.424	0.642	0.523	1628
LOCI	1.672	0.174	0.093	-0.419	0.641	0.522	1673
COF	1.692	0.177	0.108	-0.427	0.643	0.521	1683
LOOP	1.726	0.170	0.111	-0.426	0.644	0.520	1653
KNN_IN	1.715	0.172	0.109	-0.435	0.640	0.519	1671
ZSCORE	1.714	0.168	0.109	-0.422	0.651	0.509	1677
MAHA	1.748	0.167	0.096	-0.415	0.645	0.501	1935
ABOD	1.817	0.159	0.093	-0.401	0.651	0.491	1958
SOD	1.771	0.162	0.048	-0.276	0.645	0.490	2439
DB	1.766	0.162	0.116	-0.415	0.666	0.488	1677
INFLO	1.798	0.162	0.061	-0.377	0.646	0.481	2034
RNDM	1.763	0.156	0.041	-0.280	0.641	0.471	2718

Table 46: K-means clustering results for ForestFires with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.629	0.184	0.119	-0.433	0.630	0.539	1484
LOF	1.639	0.180	0.120	-0.430	0.637	0.536	1420
LOOP	1.638	0.178	0.117	-0.430	0.635	0.536	1441
KNN_IN	1.666	0.174	0.124	-0.425	0.639	0.532	1475
FBOD	1.664	0.175	0.131	-0.427	0.639	0.530	1450
LOCI	1.689	0.173	0.114	-0.427	0.638	0.529	1482
KNN_AGG	1.639	0.177	0.117	-0.431	0.651	0.519	1442
ZSCORE	1.730	0.169	0.112	-0.422	0.651	0.512	1466
SOD	1.655	0.175	0.059	-0.336	0.632	0.506	1998
MAHA	1.767	0.167	0.103	-0.413	0.649	0.497	1805
INFLO	1.757	0.165	0.072	-0.373	0.645	0.490	1844
DB	1.729	0.168	0.117	-0.423	0.669	0.489	1454
ABOD	1.808	0.153	0.088	-0.393	0.655	0.487	1850
RNDM	1.665	0.165	0.047	-0.281	0.637	0.485	2600

Table 47: K-means clustering results for ForestFires with 15% removed observations

9.4.4. GesturePhase

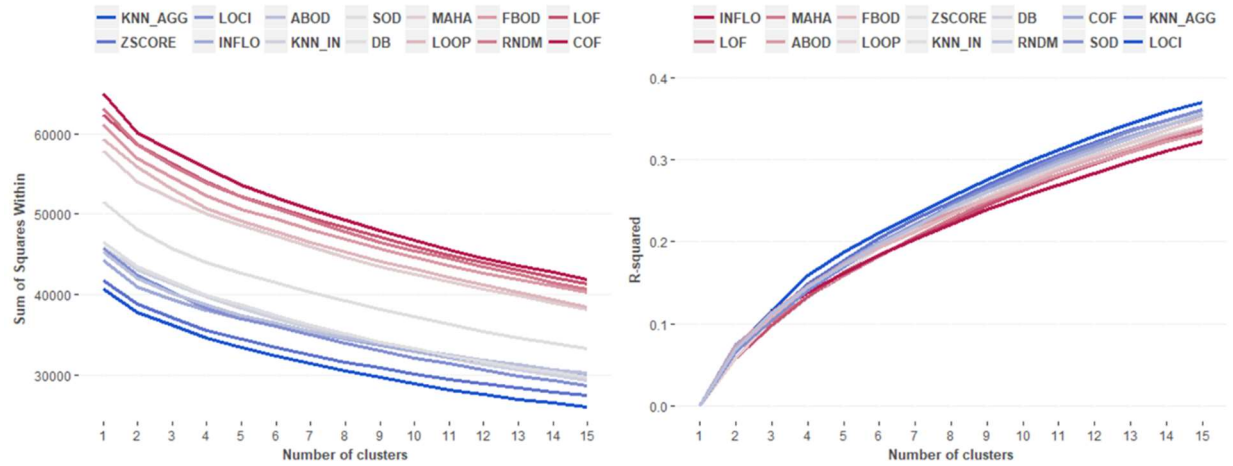


Figure 52: Hierarchical clustering for GesturePhase with 5% of observations removed

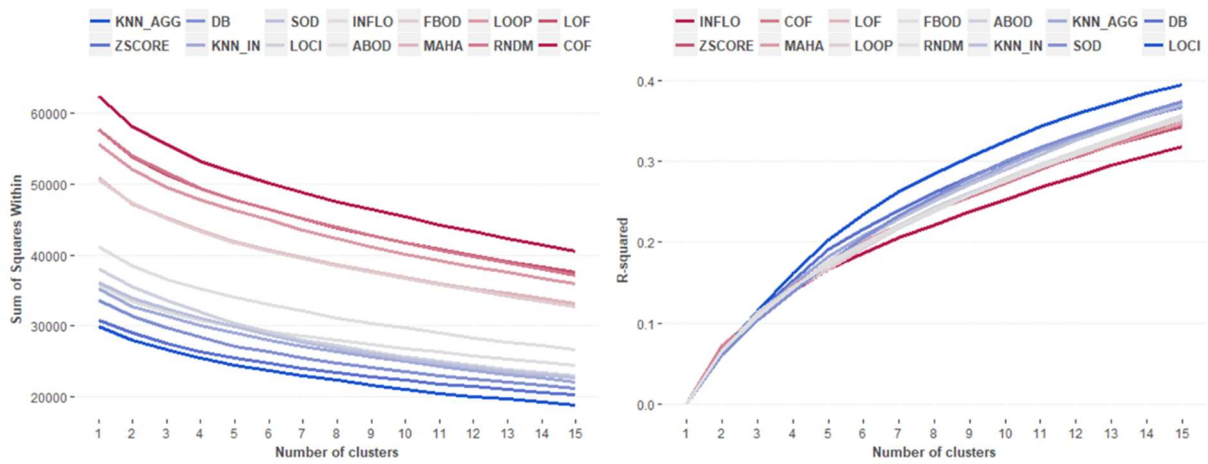


Figure 53: Hierarchical clustering for GesturePhase with 10% of observations removed

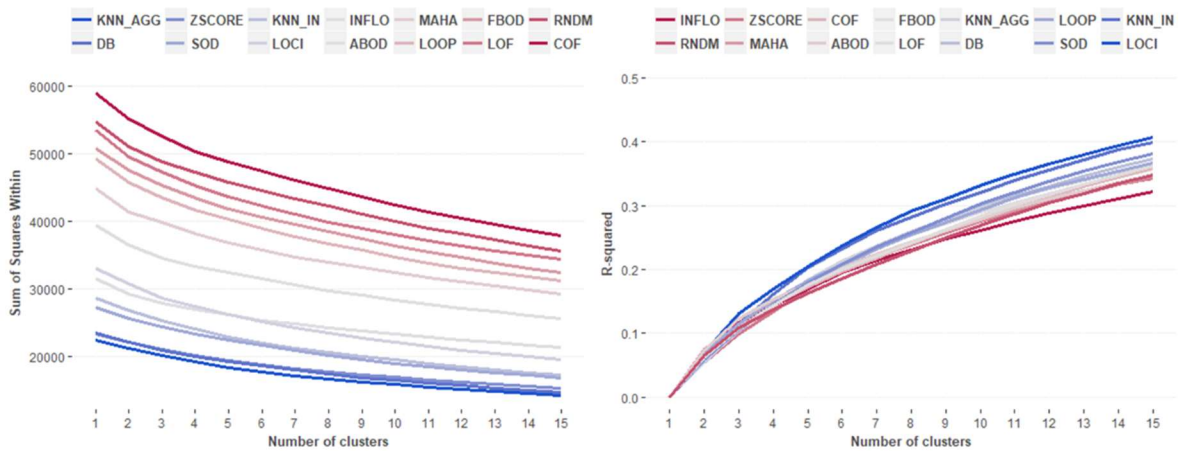


Figure 54: Hierarchical clustering results for GesturePhase with 15% removed observations

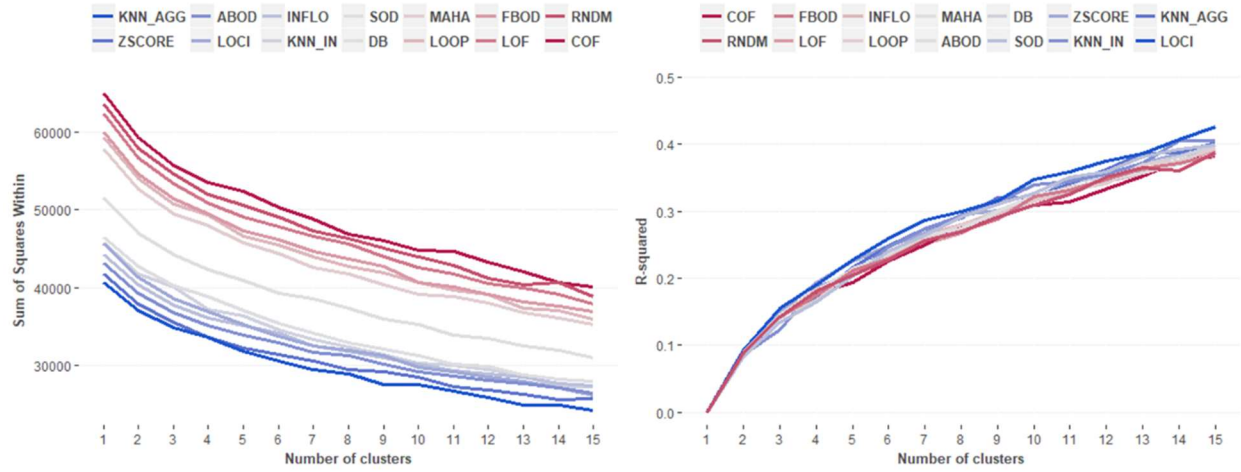


Figure 55: K-means clustering for GesturePhase with 5% of observations removed

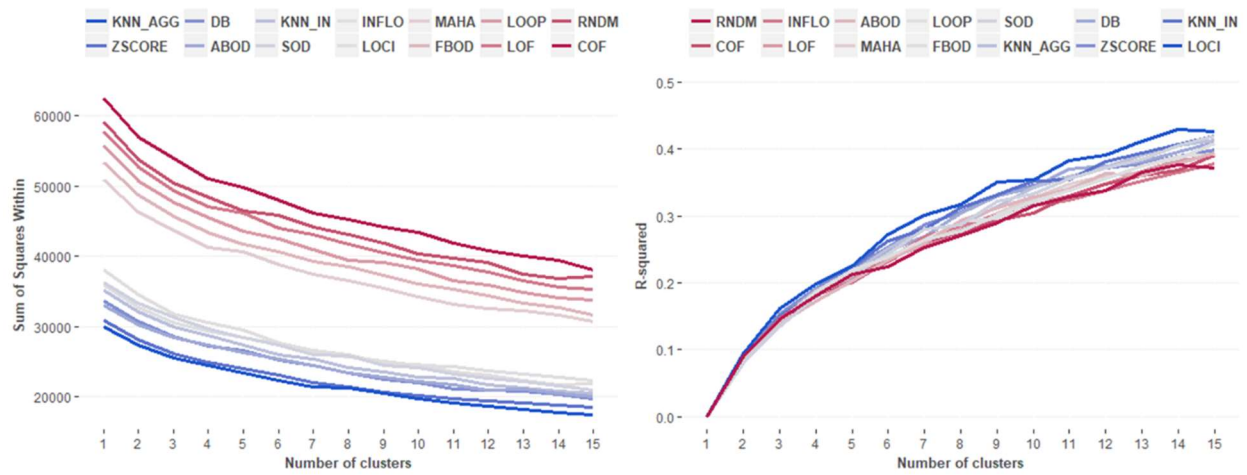


Figure 56: K-means clustering for GesturePhase with 10% of observations removed

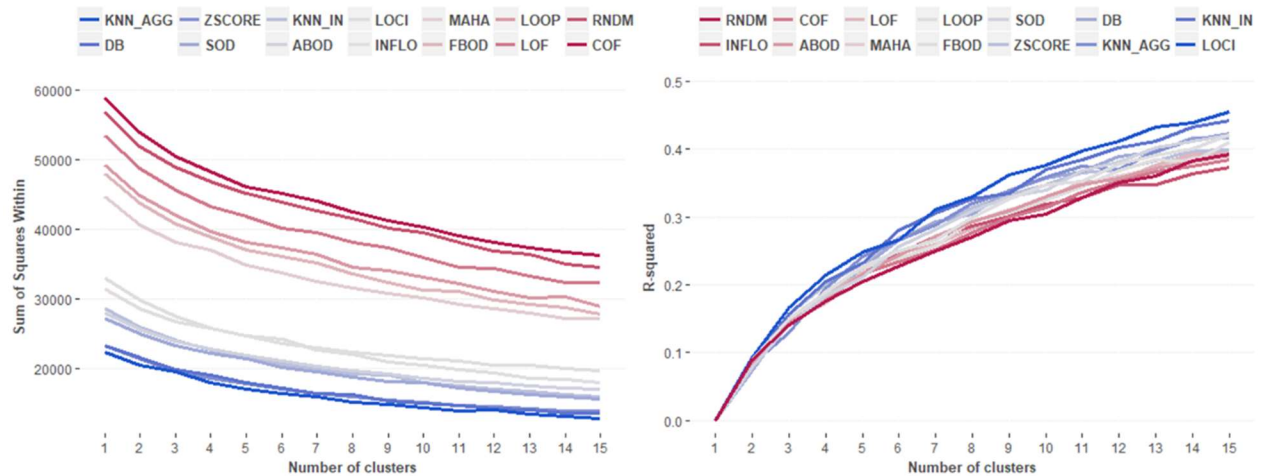


Figure 57: K-means clustering for GesturePhase with 15% of observations removed

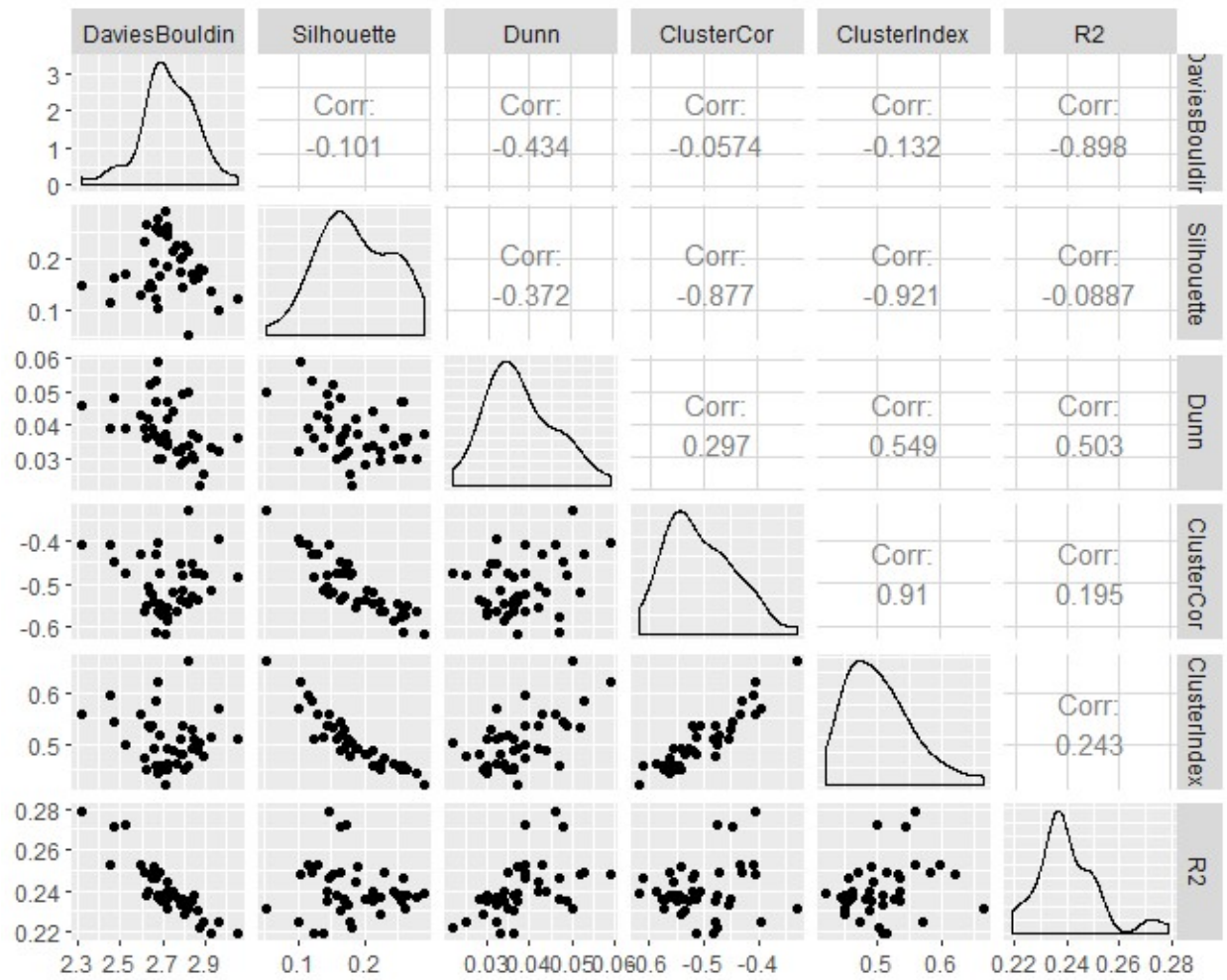


Figure 58: Correlation matrix plot for clustering evaluation metrics for GesturePhase with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	2.660	0.190	0.037	-0.541	0.491	0.252	34161
KNN_AGG	2.718	0.185	0.042	-0.556	0.493	0.244	30731
SOD	2.718	0.242	0.034	-0.586	0.462	0.240	35354
COF	2.711	0.288	0.037	-0.618	0.421	0.239	49430
RNDM	2.618	0.264	0.036	-0.552	0.449	0.238	48081
DB	2.682	0.248	0.030	-0.546	0.457	0.238	39290
KNN_IN	2.837	0.169	0.031	-0.453	0.528	0.236	34999
ZSCORE	2.833	0.168	0.037	-0.528	0.492	0.235	31911
LOOP	2.783	0.175	0.032	-0.455	0.510	0.233	45500
FBOD	2.721	0.261	0.035	-0.572	0.451	0.231	47038
ABOD	2.858	0.163	0.036	-0.537	0.488	0.230	34892
MAHA	2.888	0.178	0.025	-0.481	0.475	0.225	44834
LOF	2.873	0.180	0.022	-0.476	0.502	0.222	48442
INFLO	2.922	0.139	0.033	-0.515	0.515	0.219	34523

Table 48: Hierarchical clustering results for GesturePhase with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	2.523	0.171	0.039	-0.477	0.499	0.272	27679
DB	2.453	0.116	0.039	-0.409	0.598	0.253	25067
SOD	2.686	0.165	0.035	-0.474	0.516	0.249	27143
KNN_AGG	2.640	0.151	0.052	-0.519	0.532	0.248	22462
KNN_IN	2.651	0.146	0.039	-0.522	0.535	0.246	26498
ABOD	2.846	0.158	0.030	-0.476	0.509	0.238	31389
RNDM	2.671	0.276	0.030	-0.564	0.444	0.237	44091
FBOD	2.783	0.200	0.028	-0.536	0.481	0.236	38688
LOOP	2.790	0.204	0.033	-0.516	0.481	0.236	42546
LOF	2.719	0.254	0.047	-0.578	0.457	0.236	44151
MAHA	2.817	0.213	0.034	-0.548	0.457	0.235	38873
COF	2.697	0.254	0.030	-0.573	0.450	0.235	47775
ZSCORE	2.793	0.143	0.049	-0.480	0.537	0.235	23599
INFLO	3.048	0.124	0.036	-0.483	0.510	0.219	28038

Table 49: Hierarchical clustering results for GesturePhase with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
LOCI	2.314	0.147	0.046	-0.407	0.558	0.279	23766
KNN_IN	2.468	0.163	0.048	-0.449	0.543	0.271	20870
SOD	2.597	0.130	0.043	-0.433	0.558	0.253	20345
LOOP	2.616	0.230	0.039	-0.564	0.471	0.249	36909
DB	2.666	0.121	0.053	-0.431	0.584	0.249	17583
KNN_AGG	2.676	0.105	0.059	-0.405	0.621	0.248	16887
LOF	2.668	0.258	0.047	-0.612	0.457	0.246	40255
FBOD	2.749	0.212	0.044	-0.521	0.487	0.240	38520
ABOD	2.631	0.143	0.042	-0.508	0.536	0.240	29900
COF	2.678	0.257	0.036	-0.564	0.455	0.238	44883
MAHA	2.763	0.224	0.032	-0.564	0.462	0.237	34122
ZSCORE	2.819	0.054	0.050	-0.331	0.664	0.231	17915
RNDM	2.796	0.223	0.029	-0.543	0.450	0.229	42140
INFLO	2.957	0.101	0.032	-0.397	0.572	0.225	24406

Table 50: Hierarchical clustering results for GesturePhase with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
LOCI	2.384	0.200	0.040	-0.534	0.462	0.295	32179
KNN_AGG	2.513	0.141	0.043	-0.484	0.496	0.283	29160
KNN_IN	2.441	0.174	0.037	-0.502	0.476	0.282	32873
ZSCORE	2.501	0.174	0.037	-0.513	0.478	0.281	30006
SOD	2.510	0.157	0.030	-0.472	0.483	0.279	33557
DB	2.515	0.182	0.027	-0.477	0.461	0.274	37404
ABOD	2.531	0.170	0.035	-0.503	0.478	0.273	31315
MAHA	2.564	0.186	0.033	-0.507	0.448	0.272	42086
LOOP	2.572	0.209	0.037	-0.542	0.438	0.272	43175
INFLO	2.550	0.168	0.038	-0.505	0.474	0.272	32190
LOF	2.575	0.195	0.034	-0.518	0.441	0.271	45431
FBOD	2.538	0.200	0.034	-0.532	0.443	0.270	43761
RNDM	2.597	0.159	0.024	-0.470	0.455	0.269	46479
COF	2.624	0.166	0.026	-0.481	0.451	0.266	47723

Table 51: K-means clustering results for GesturePhase with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	2.332	0.173	0.043	-0.500	0.471	0.308	26301
KNN_IN	2.381	0.179	0.043	-0.516	0.485	0.296	24760
ZSCORE	2.452	0.160	0.044	-0.502	0.500	0.291	21849
DB	2.421	0.175	0.032	-0.489	0.492	0.290	23829
KNN_AGG	2.407	0.185	0.044	-0.537	0.499	0.289	21242
SOD	2.449	0.172	0.033	-0.483	0.485	0.286	25786
FBOD	2.496	0.208	0.044	-0.550	0.449	0.279	38442
LOOP	2.511	0.195	0.036	-0.531	0.453	0.278	40168
MAHA	2.546	0.182	0.037	-0.501	0.456	0.277	36760
ABOD	2.513	0.166	0.034	-0.498	0.490	0.276	23903
LOF	2.568	0.160	0.034	-0.492	0.462	0.273	41968
INFLO	2.550	0.172	0.037	-0.496	0.478	0.271	26141
COF	2.587	0.179	0.028	-0.496	0.452	0.270	45618
RNDM	2.607	0.181	0.028	-0.497	0.444	0.269	43157

Table 52: K-means clustering results for GesturePhase with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOCI	2.255	0.194	0.044	-0.511	0.466	0.322	22366
KNN_IN	2.259	0.178	0.045	-0.505	0.490	0.313	19688
KNN_AGG	2.359	0.148	0.052	-0.475	0.541	0.300	15712
DB	2.383	0.159	0.039	-0.473	0.534	0.299	16411
ZSCORE	2.402	0.156	0.046	-0.487	0.521	0.295	16415
SOD	2.350	0.169	0.032	-0.472	0.508	0.294	19212
FBOD	2.428	0.198	0.036	-0.539	0.457	0.292	34009
LOOP	2.491	0.187	0.036	-0.526	0.460	0.283	35241
MAHA	2.500	0.162	0.036	-0.484	0.467	0.281	32159
LOF	2.502	0.186	0.038	-0.522	0.461	0.280	38404
ABOD	2.466	0.145	0.038	-0.467	0.513	0.280	20118
COF	2.518	0.208	0.030	-0.528	0.446	0.273	42842
INFLO	2.546	0.165	0.035	-0.479	0.481	0.273	22908
RNDM	2.571	0.192	0.028	-0.499	0.444	0.270	41539

Table 53: K-means clustering results for GesturePhase with 15% removed observations

9.5. FIGURES FOR COHERENT RESULTS

9.5.1. PimaIndians

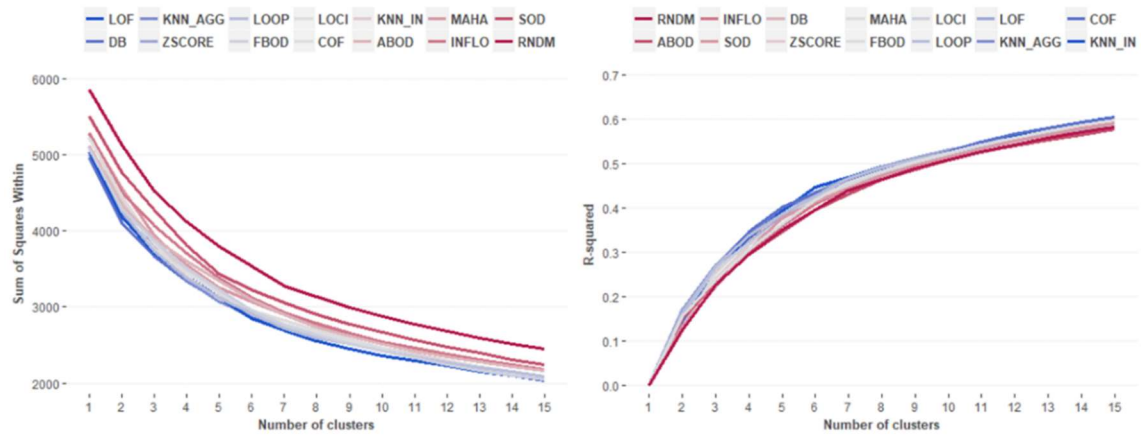


Figure 59: Hierarchical clustering for PimaIndians with 5% removed observations

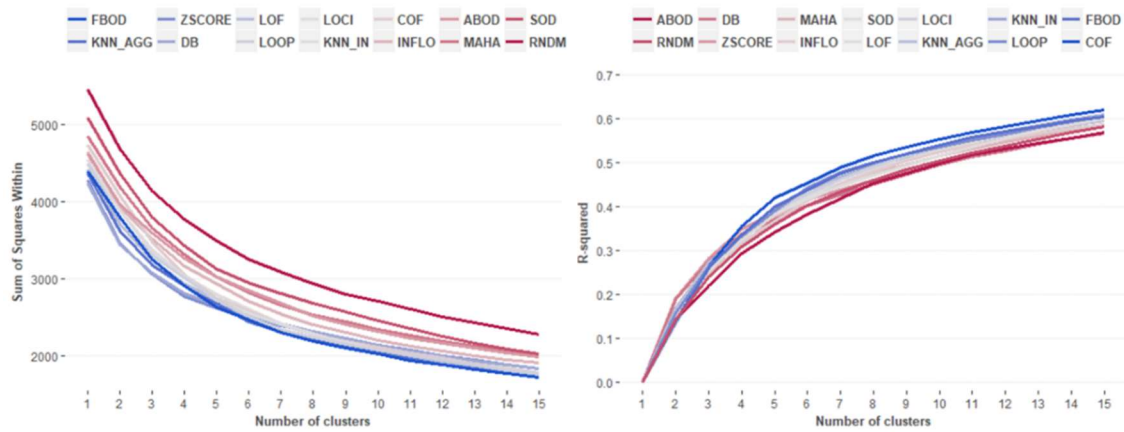


Figure 60: Hierarchical clustering for PimaIndians with 10% removed observations

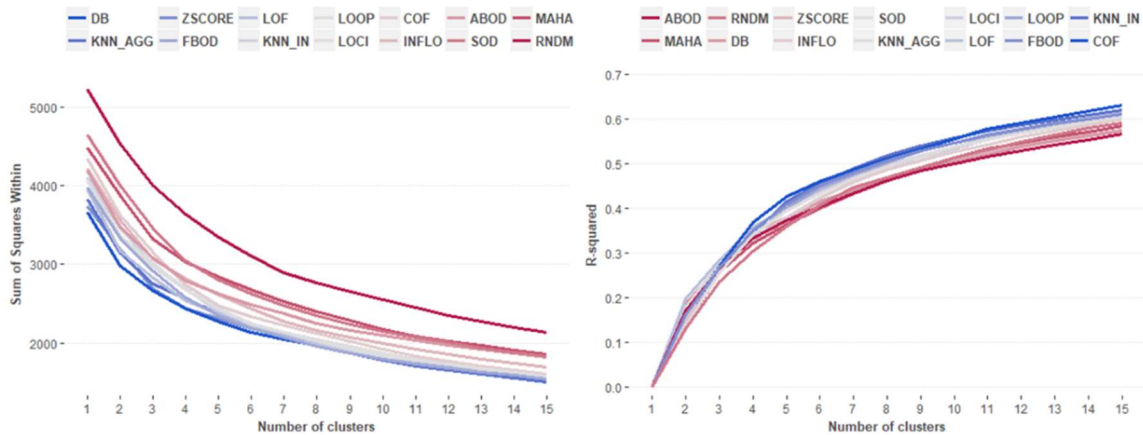


Figure 61: Hierarchical clustering for PimaIndians with 15% removed observations

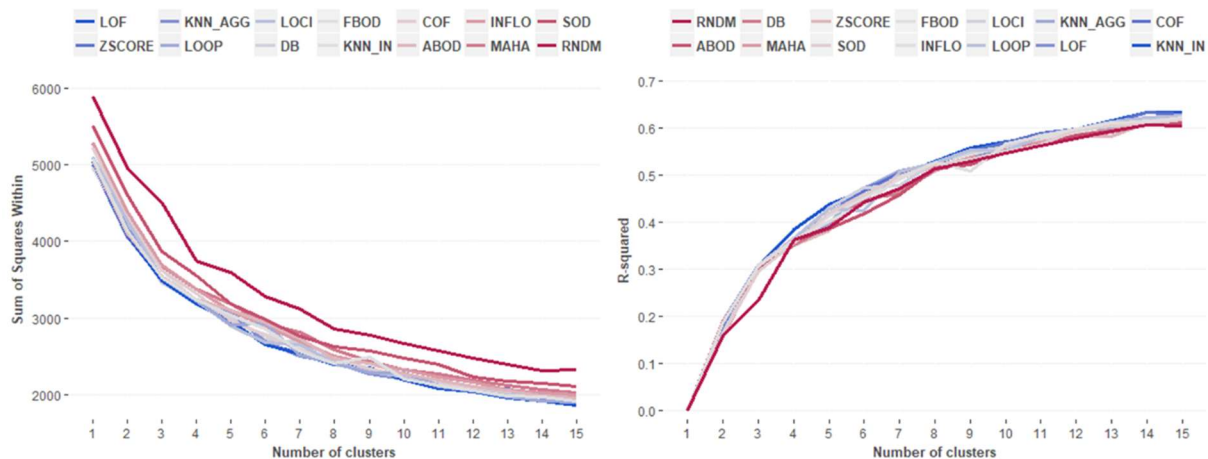


Figure 62: K-means clustering for PimaIndians with 5% removed observations

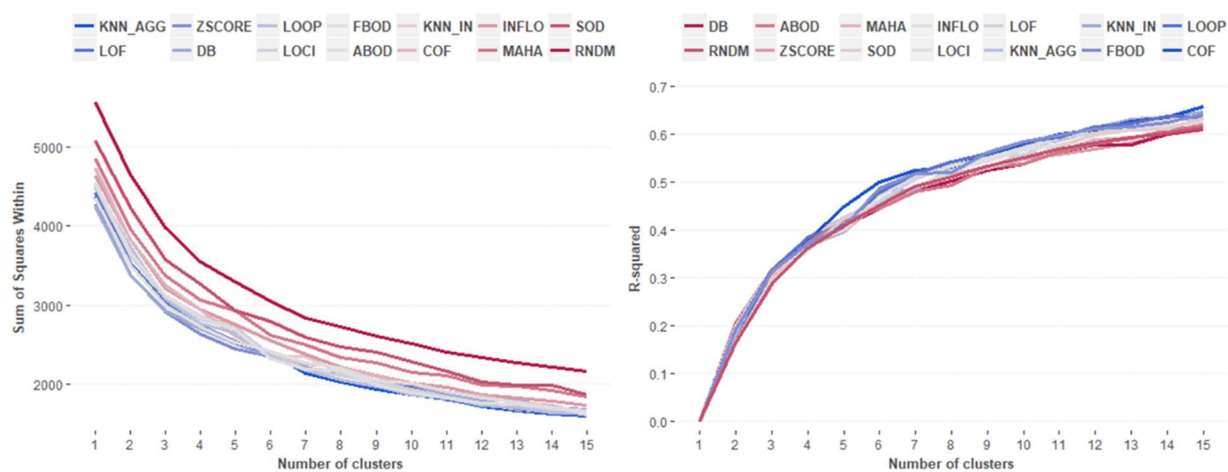


Figure 63: K-means clustering for PimaIndians with 10% removed observations

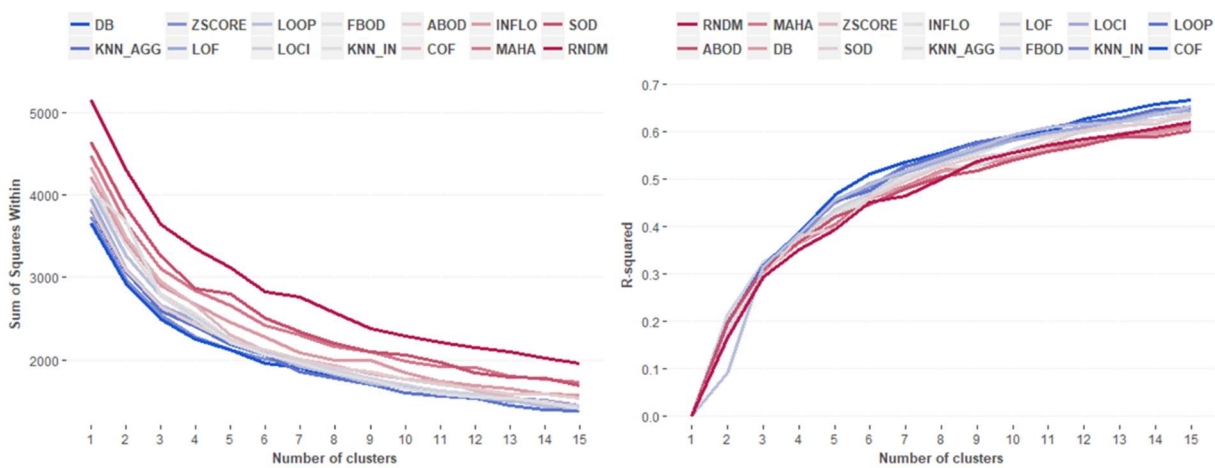


Figure 64: K-means clustering for PimaIndians with 15% removed observations

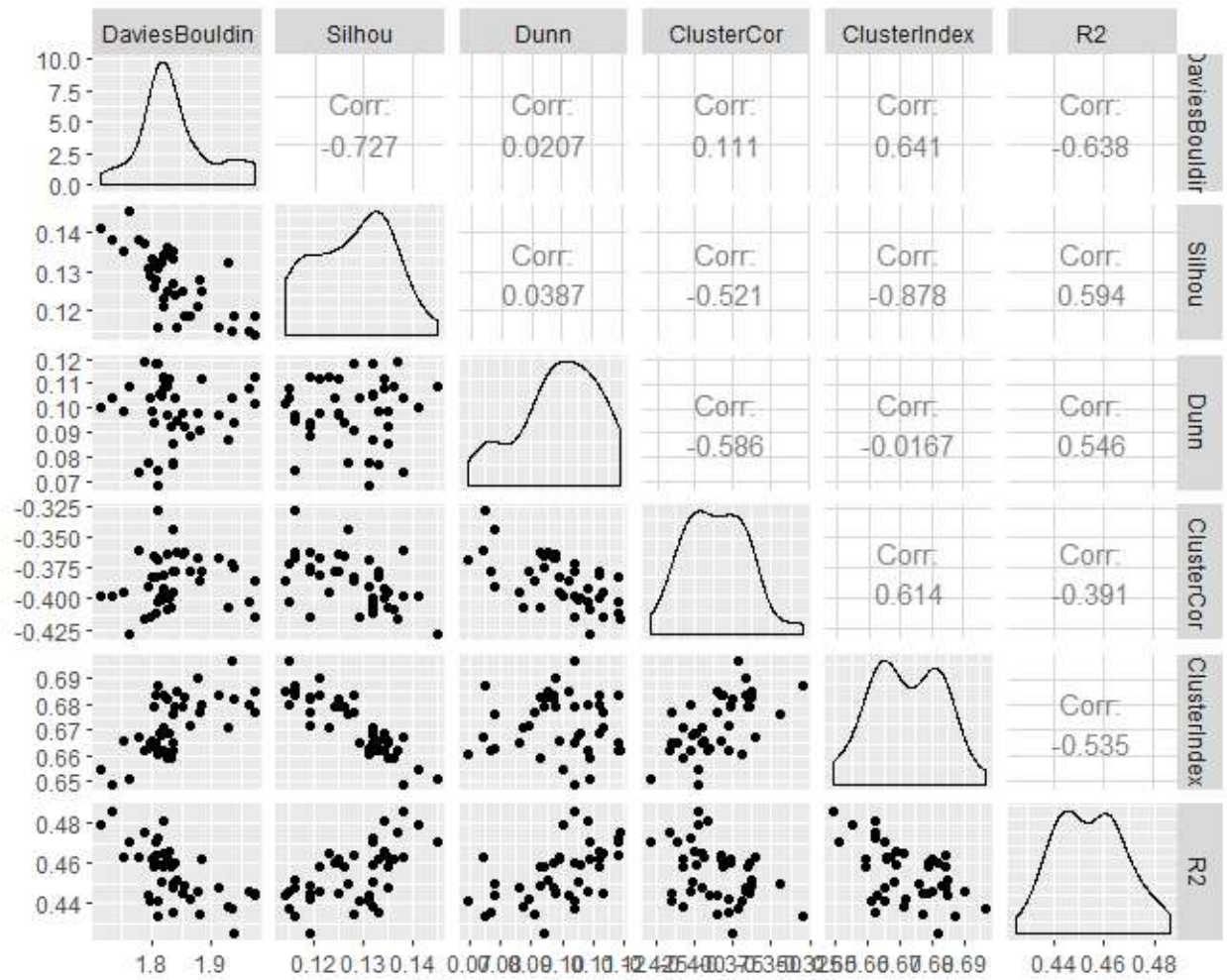


Figure 65: Correlation matrix plot for clustering evaluation metrics for PimaIndians with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	1.752	0.135	0.099	-0.394	0.666	0.463	2808
COF	1.800	0.133	0.099	-0.383	0.666	0.462	2807
KNN_AGG	1.827	0.125	0.097	-0.364	0.682	0.460	2753
LOF	1.839	0.124	0.104	-0.377	0.679	0.460	2723
LOOP	1.832	0.135	0.093	-0.407	0.659	0.458	2764
LOCI	1.843	0.116	0.095	-0.363	0.685	0.452	2796
FBOD	1.815	0.132	0.105	-0.401	0.668	0.451	2789
MAHA	1.836	0.135	0.086	-0.395	0.665	0.448	2911
ZSCORE	1.910	0.116	0.097	-0.367	0.684	0.448	2763
DB	1.876	0.121	0.098	-0.367	0.690	0.446	2750
SOD	1.793	0.131	0.078	-0.390	0.663	0.444	3063
INFLO	1.864	0.119	0.089	-0.378	0.672	0.442	2946
ABOD	1.881	0.128	0.091	-0.385	0.677	0.435	2890
RNDM	1.809	0.116	0.075	-0.329	0.687	0.434	3313

Table 54: Hierarchical clustering results for PimaIndians with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.716	0.141	0.100	-0.398	0.655	0.479	2468
FBOD	1.828	0.134	0.112	-0.400	0.669	0.466	2349
LOOP	1.818	0.123	0.113	-0.395	0.671	0.465	2404
KNN_IN	1.826	0.136	0.109	-0.408	0.659	0.462	2441
KNN_AGG	1.884	0.125	0.112	-0.377	0.680	0.462	2350
LOCI	1.812	0.132	0.106	-0.399	0.669	0.459	2431
LOF	1.820	0.121	0.112	-0.381	0.684	0.458	2395
SOD	1.836	0.127	0.078	-0.344	0.676	0.450	2796
INFLO	1.853	0.119	0.093	-0.363	0.683	0.449	2558
MAHA	1.850	0.125	0.098	-0.364	0.679	0.445	2691
ZSCORE	1.972	0.114	0.102	-0.385	0.685	0.444	2377
DB	1.934	0.115	0.104	-0.371	0.697	0.438	2386
RNDM	1.835	0.133	0.077	-0.378	0.662	0.436	3078
ABOD	1.938	0.119	0.094	-0.375	0.682	0.425	2657

Table 55: Hierarchical clustering results for PimaIndians with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.733	0.138	0.104	-0.398	0.649	0.486	2228
KNN_IN	1.818	0.134	0.108	-0.391	0.662	0.481	2128
FBOD	1.787	0.137	0.119	-0.416	0.662	0.475	2088
LOOP	1.809	0.132	0.118	-0.403	0.662	0.472	2146
LOF	1.806	0.132	0.118	-0.411	0.665	0.471	2089
LOCI	1.763	0.145	0.109	-0.429	0.651	0.471	2148
KNN_AGG	1.806	0.128	0.118	-0.383	0.684	0.464	2051
SOD	1.779	0.138	0.074	-0.361	0.667	0.463	2490
INFLO	1.804	0.126	0.094	-0.366	0.679	0.458	2287
ZSCORE	1.962	0.115	0.108	-0.403	0.680	0.446	2067
DB	1.971	0.119	0.113	-0.415	0.677	0.445	2032
RNDM	1.810	0.131	0.069	-0.369	0.661	0.441	2917
MAHA	1.798	0.129	0.104	-0.415	0.665	0.441	2499
ABOD	1.926	0.132	0.087	-0.407	0.671	0.439	2352

Table 56: Hierarchical clustering results for PimaIndians with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	1.672	0.165	0.076	-0.413	0.629	0.499	2622
COF	1.674	0.166	0.070	-0.407	0.634	0.497	2624
LOF	1.692	0.168	0.078	-0.426	0.636	0.494	2549
KNN_AGG	1.720	0.159	0.068	-0.395	0.644	0.493	2586
LOOP	1.680	0.167	0.074	-0.423	0.635	0.493	2587
LOCI	1.702	0.165	0.069	-0.418	0.637	0.492	2588
INFLO	1.697	0.162	0.066	-0.405	0.634	0.488	2704
FBOD	1.712	0.160	0.077	-0.414	0.640	0.487	2609
SOD	1.677	0.164	0.059	-0.397	0.638	0.485	2835
ZSCORE	1.745	0.155	0.072	-0.411	0.645	0.483	2585
MAHA	1.723	0.159	0.068	-0.390	0.648	0.480	2742
DB	1.781	0.153	0.075	-0.399	0.656	0.478	2592
ABOD	1.748	0.158	0.068	-0.400	0.650	0.476	2650
RNDM	1.721	0.157	0.056	-0.387	0.637	0.471	3115

Table 57: K-means clustering results for PimaIndians with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.668	0.164	0.073	-0.410	0.627	0.511	2315
LOOP	1.664	0.170	0.082	-0.428	0.634	0.505	2222
FBOD	1.668	0.169	0.082	-0.428	0.633	0.504	2250
KNN_IN	1.685	0.163	0.083	-0.413	0.638	0.503	2259
KNN_AGG	1.675	0.170	0.079	-0.425	0.639	0.502	2176
LOF	1.667	0.170	0.088	-0.434	0.635	0.502	2203
LOCI	1.679	0.169	0.083	-0.426	0.635	0.501	2245
INFLO	1.683	0.164	0.072	-0.408	0.640	0.490	2365
SOD	1.710	0.161	0.060	-0.390	0.637	0.487	2610
MAHA	1.685	0.165	0.070	-0.404	0.643	0.485	2500
ZSCORE	1.783	0.154	0.078	-0.431	0.643	0.484	2206
ABOD	1.713	0.159	0.078	-0.421	0.649	0.482	2252
RNDM	1.700	0.163	0.059	-0.395	0.630	0.480	2895
DB	1.736	0.159	0.077	-0.425	0.652	0.479	2211

Table 58: K-means clustering results for PimaIndians with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.624	0.178	0.079	-0.423	0.620	0.524	2064
LOOP	1.626	0.175	0.086	-0.434	0.628	0.516	1965
KNN_IN	1.656	0.169	0.092	-0.422	0.633	0.512	2003
LOCI	1.664	0.174	0.088	-0.438	0.628	0.511	1986
FBOD	1.600	0.182	0.110	-0.437	0.620	0.507	2000
LOF	1.656	0.169	0.091	-0.444	0.631	0.507	1948
KNN_AGG	1.692	0.167	0.083	-0.428	0.643	0.504	1897
INFLO	1.683	0.165	0.066	-0.409	0.638	0.493	2135
SOD	1.657	0.168	0.055	-0.396	0.631	0.491	2362
ZSCORE	1.759	0.156	0.087	-0.435	0.646	0.487	1914
DB	1.756	0.156	0.089	-0.439	0.650	0.485	1884
MAHA	1.712	0.159	0.072	-0.403	0.644	0.483	2310
ABOD	1.779	0.153	0.076	-0.421	0.650	0.479	2008
RNDM	1.706	0.159	0.054	-0.391	0.633	0.478	2691

Table 59: K-means clustering results for PimaIndians with 15% removed observations

9.5.2. Ionosphere

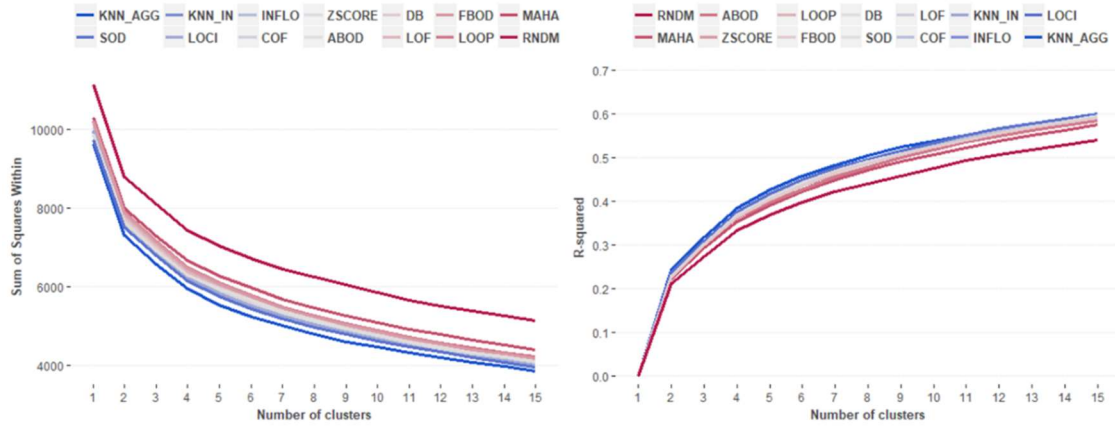


Figure 66: Hierarchical clustering for Ionosphere with 5% of observations removed

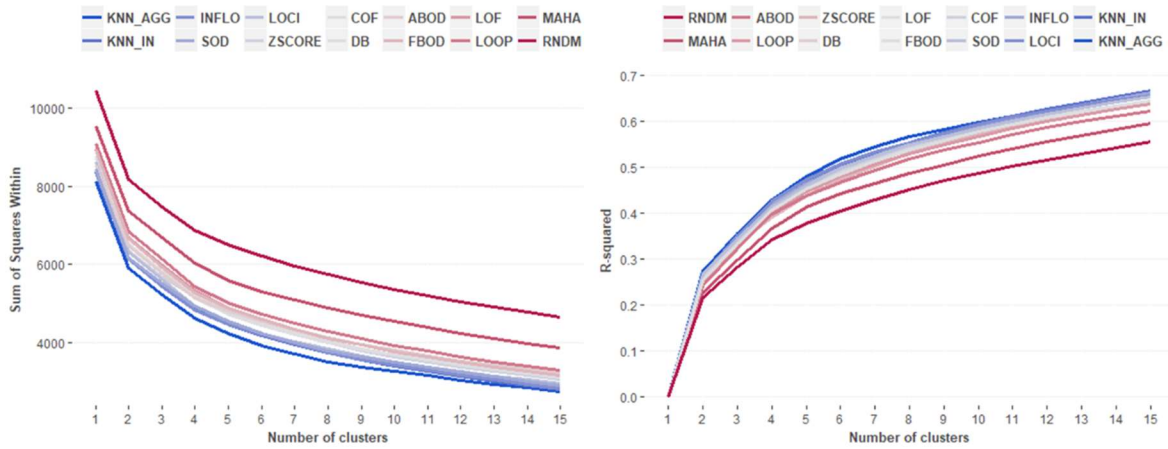


Figure 67: Hierarchical clustering for Ionosphere with 10% of observations removed

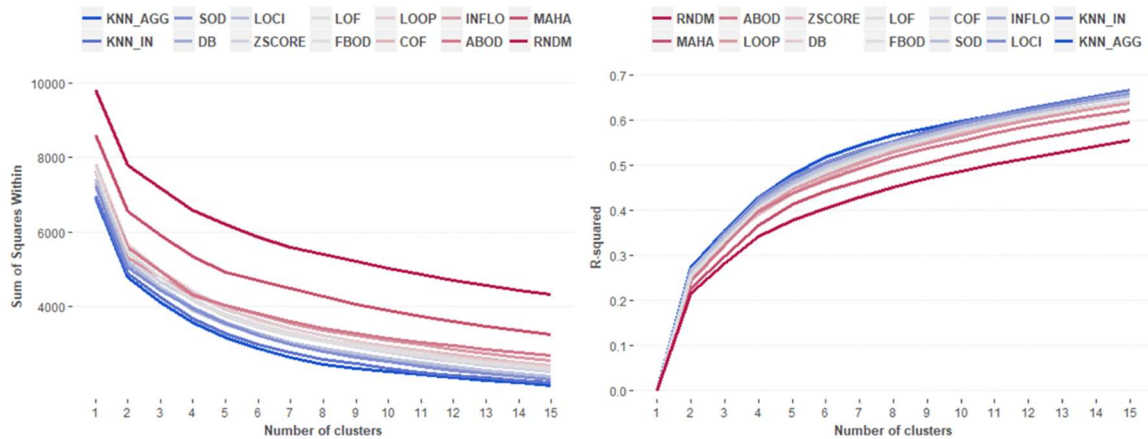


Figure 68: Hierarchical clustering for Ionosphere with 15% of observations removed

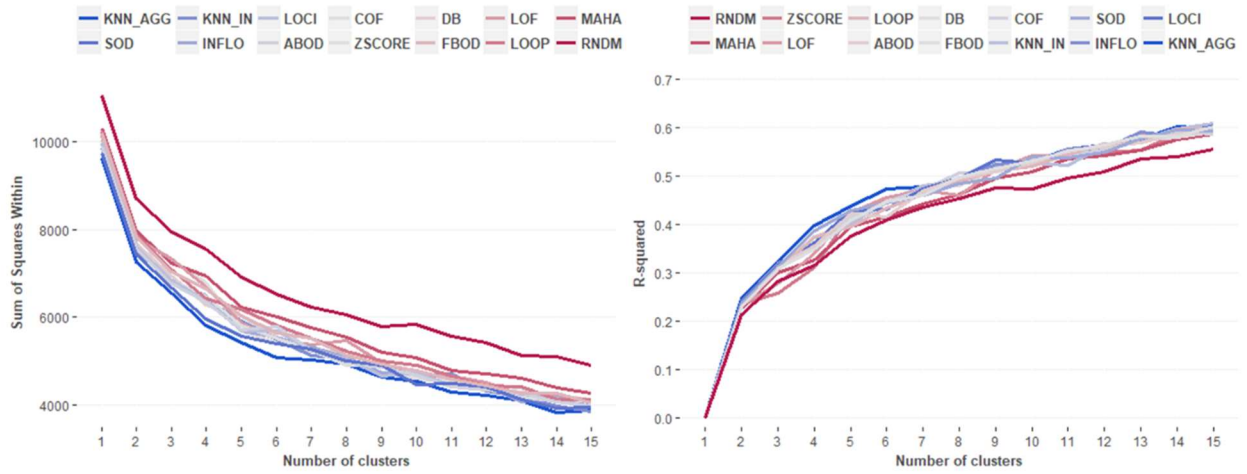


Figure 69: K-means clustering for Ionosphere with 5% of observations removed

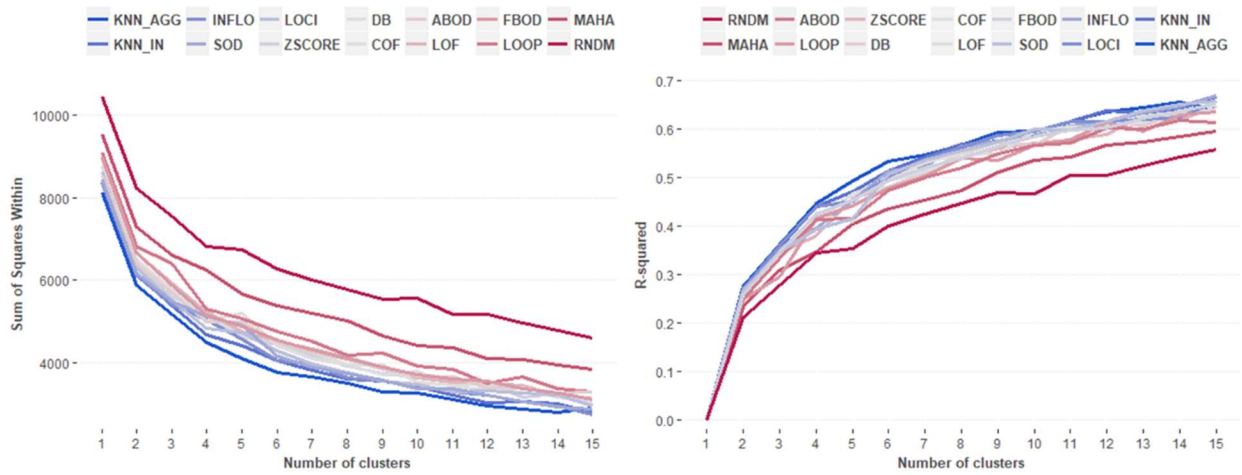


Figure 70: K-means clustering for Ionosphere with 10% of observations removed

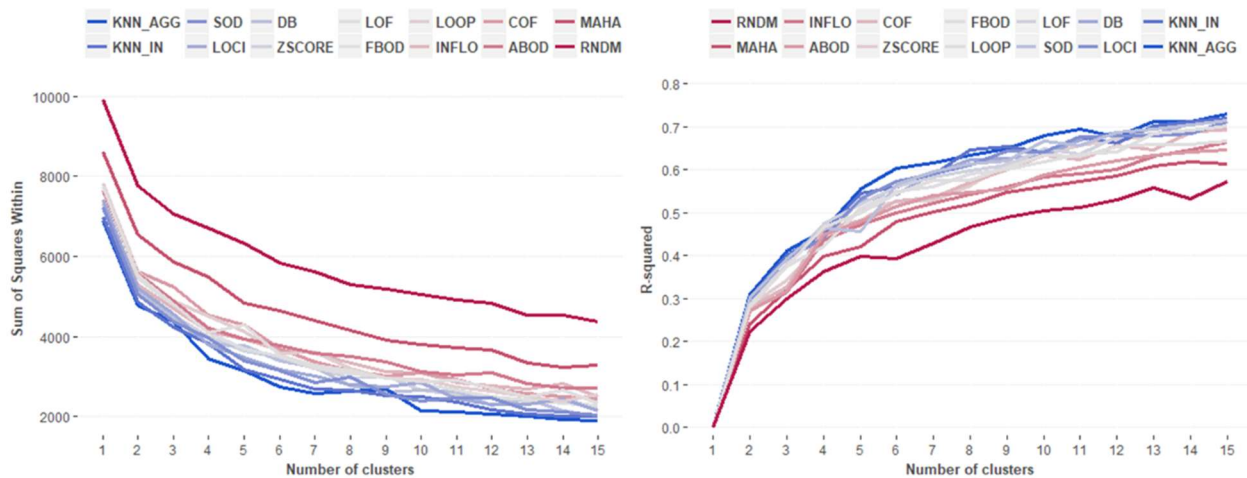


Figure 71: K-means clustering for Ionosphere with 15% of observations removed

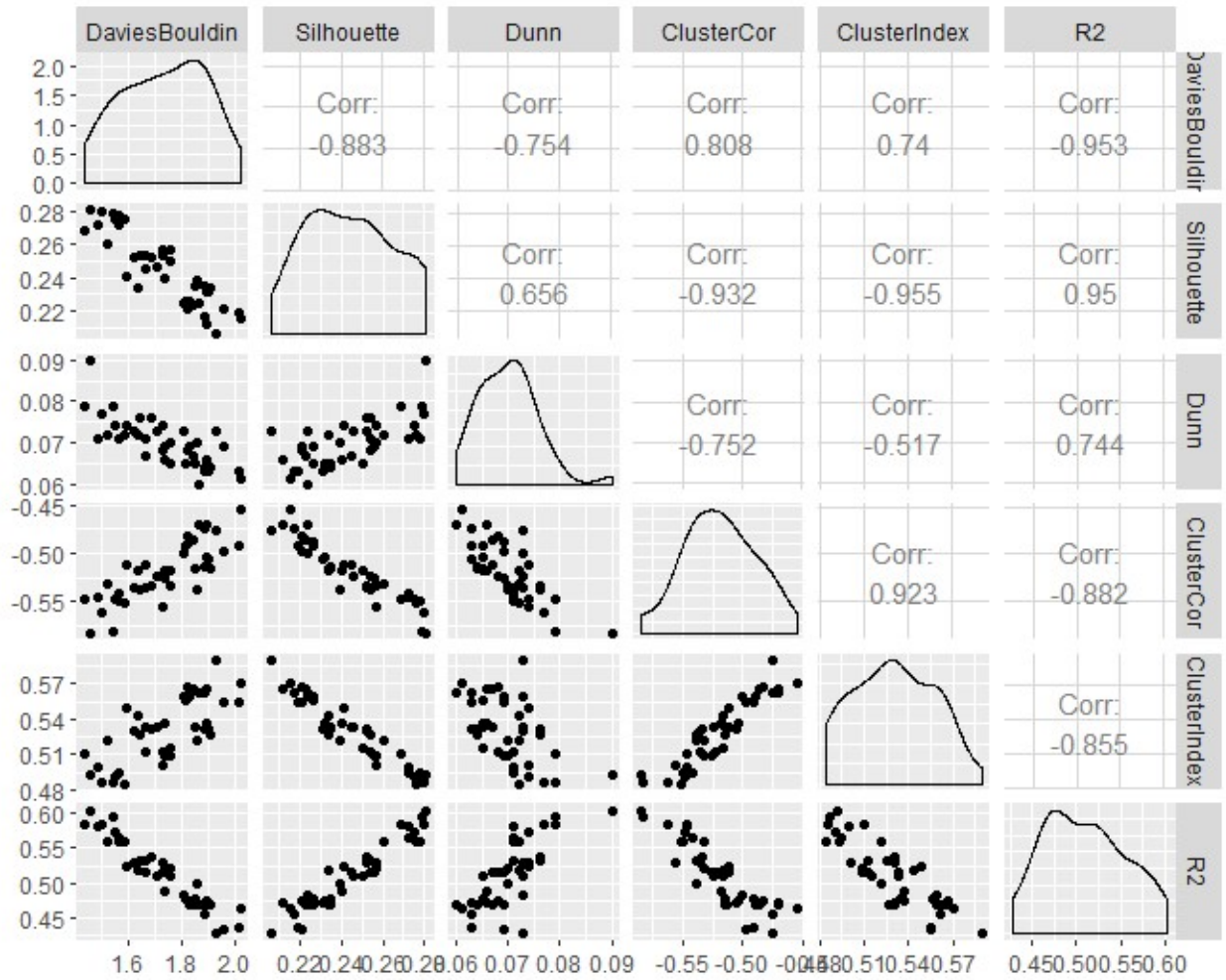


Figure 72: Correlation matrix plot for clustering evaluation metrics for Ionosphere with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.804	0.224	0.073	-0.500	0.559	0.483	4986
LOCI	1.848	0.235	0.065	-0.515	0.534	0.478	5211
INFLO	1.810	0.226	0.065	-0.492	0.556	0.477	5212
KNN_IN	1.817	0.221	0.068	-0.483	0.567	0.475	5191
COF	1.894	0.232	0.065	-0.504	0.537	0.475	5231
LOF	1.900	0.231	0.063	-0.506	0.531	0.472	5360
SOD	1.891	0.212	0.066	-0.471	0.565	0.471	5155
DB	1.827	0.226	0.069	-0.489	0.559	0.470	5326
FBOD	1.889	0.235	0.064	-0.513	0.531	0.469	5424
LOOP	1.907	0.234	0.064	-0.515	0.527	0.468	5451
ZSCORE	1.841	0.223	0.067	-0.486	0.565	0.468	5248
ABOD	2.024	0.215	0.061	-0.455	0.571	0.463	5311
MAHA	1.889	0.217	0.063	-0.475	0.562	0.454	5633
RNDM	1.929	0.206	0.073	-0.476	0.590	0.427	6396

Table 60: Hierarchical clustering results for Ionosphere with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.687	0.252	0.076	-0.534	0.531	0.538	3748
KNN_IN	1.644	0.254	0.076	-0.538	0.526	0.532	3915
LOCI	1.664	0.254	0.071	-0.536	0.512	0.532	4036
INFLO	1.622	0.252	0.073	-0.535	0.530	0.530	3932
SOD	1.589	0.241	0.074	-0.511	0.549	0.525	3999
COF	1.756	0.257	0.070	-0.533	0.509	0.520	4211
FBOD	1.733	0.256	0.069	-0.526	0.509	0.516	4315
LOF	1.727	0.254	0.068	-0.523	0.513	0.516	4331
DB	1.664	0.245	0.067	-0.511	0.533	0.515	4238
ZSCORE	1.704	0.246	0.073	-0.524	0.533	0.510	4182
LOOP	1.754	0.250	0.065	-0.517	0.515	0.509	4468
ABOD	1.856	0.239	0.070	-0.538	0.522	0.498	4292
MAHA	1.862	0.224	0.060	-0.471	0.562	0.470	5058
RNDM	2.014	0.219	0.063	-0.492	0.555	0.436	5886

Table 61: Hierarchical clustering results for Ionosphere with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
KNN_AGG	1.453	0.281	0.090	-0.584	0.492	0.604	2733
KNN_IN	1.541	0.279	0.079	-0.582	0.487	0.594	2832
LOCI	1.498	0.280	0.077	-0.562	0.486	0.584	3081
SOD	1.435	0.269	0.079	-0.547	0.510	0.583	3008
DB	1.480	0.272	0.071	-0.546	0.500	0.580	3068
LOF	1.546	0.275	0.074	-0.548	0.491	0.572	3281
FBOD	1.560	0.272	0.071	-0.541	0.495	0.565	3326
LOOP	1.584	0.276	0.072	-0.551	0.485	0.560	3435
COF	1.562	0.278	0.071	-0.549	0.494	0.560	3437
ZSCORE	1.519	0.260	0.072	-0.532	0.522	0.558	3254
ABOD	1.730	0.257	0.074	-0.556	0.501	0.529	3601
INFLO	1.635	0.234	0.072	-0.517	0.543	0.519	3511
MAHA	1.736	0.240	0.066	-0.518	0.537	0.489	4395
RNDM	1.955	0.221	0.069	-0.497	0.555	0.434	5550

Table 62: Hierarchical clustering results for Ionosphere with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R²	Avg SSW
LOCI	1.917	0.235	0.057	-0.513	0.524	0.484	5153
KNN_AGG	1.930	0.238	0.062	-0.519	0.539	0.483	4987
COF	1.919	0.240	0.055	-0.522	0.519	0.481	5168
INFLO	1.868	0.233	0.052	-0.499	0.549	0.480	5186
LOF	1.904	0.246	0.053	-0.532	0.513	0.478	5301
FBOD	1.914	0.253	0.054	-0.537	0.509	0.477	5316
SOD	1.866	0.235	0.059	-0.509	0.545	0.476	5103
KNN_IN	1.899	0.237	0.056	-0.515	0.540	0.475	5197
ABOD	2.001	0.231	0.056	-0.506	0.535	0.473	5212
DB	1.926	0.232	0.058	-0.503	0.541	0.472	5304
LOOP	1.981	0.239	0.052	-0.529	0.508	0.472	5407
ZSCORE	1.926	0.231	0.055	-0.510	0.543	0.469	5240
MAHA	1.908	0.231	0.053	-0.492	0.550	0.461	5559
RNDM	2.106	0.214	0.052	-0.491	0.551	0.434	6209

Table 63: K-means clustering results for Ionosphere with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.656	0.267	0.062	-0.557	0.516	0.543	3706
KNN_IN	1.624	0.253	0.062	-0.523	0.526	0.537	3874
LOCI	1.810	0.262	0.054	-0.548	0.494	0.532	4040
INFLO	1.692	0.250	0.065	-0.524	0.522	0.529	3934
SOD	1.705	0.239	0.056	-0.512	0.531	0.526	3996
FBOD	1.820	0.266	0.061	-0.546	0.493	0.522	4295
LOF	1.802	0.251	0.056	-0.525	0.503	0.521	4285
COF	1.836	0.258	0.059	-0.542	0.501	0.520	4213
DB	1.726	0.256	0.058	-0.528	0.515	0.519	4195
ZSCORE	1.766	0.230	0.063	-0.507	0.532	0.512	4162
LOOP	1.832	0.251	0.057	-0.528	0.507	0.506	4495
ABOD	1.917	0.245	0.059	-0.527	0.514	0.502	4252
MAHA	1.962	0.228	0.051	-0.495	0.550	0.469	5059
RNDM	2.070	0.203	0.057	-0.475	0.565	0.431	5938

Table 64: K-means clustering results for Ionosphere with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	1.555	0.269	0.068	-0.562	0.491	0.601	2753
KNN_IN	1.610	0.263	0.068	-0.547	0.492	0.591	2852
SOD	1.495	0.274	0.065	-0.544	0.503	0.585	2993
LOCI	1.631	0.273	0.059	-0.557	0.478	0.583	3090
DB	1.548	0.278	0.058	-0.556	0.494	0.577	3094
LOF	1.664	0.292	0.063	-0.582	0.467	0.573	3268
ZSCORE	1.548	0.280	0.067	-0.561	0.495	0.565	3199
FBOD	1.741	0.268	0.055	-0.544	0.478	0.561	3402
LOOP	1.650	0.280	0.056	-0.558	0.485	0.559	3444
COF	1.696	0.245	0.054	-0.502	0.508	0.551	3506
ABOD	1.770	0.260	0.059	-0.547	0.498	0.530	3592
INFLO	1.734	0.243	0.067	-0.526	0.521	0.526	3460
MAHA	1.846	0.241	0.049	-0.511	0.528	0.495	4345
RNDM	2.100	0.226	0.055	-0.510	0.555	0.438	5575

Table 65: K-means clustering results for Ionosphere with 15% removed observations

9.5.3. HeartDisease

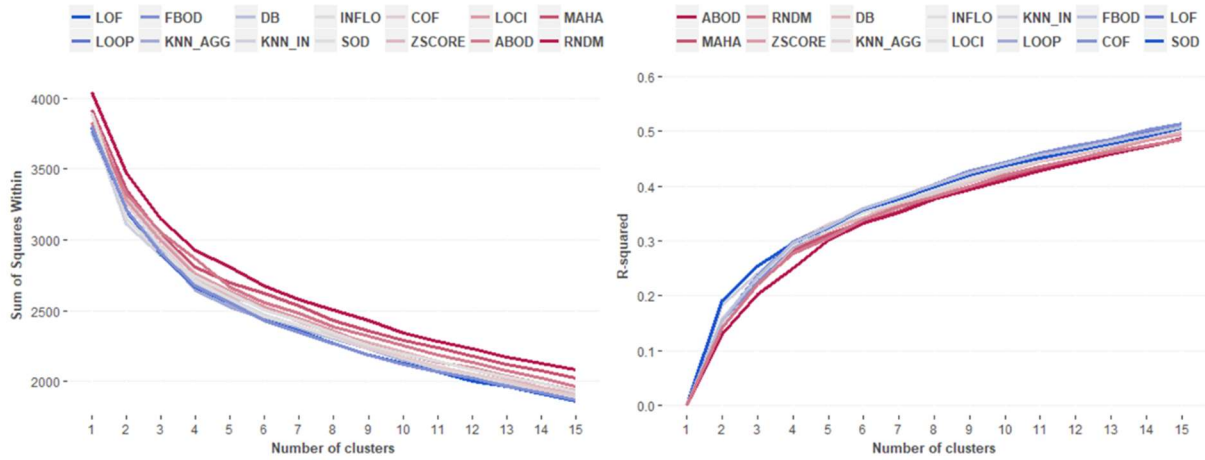


Figure 73: Hierarchical clustering for HeartDisease with 5% of observations removed

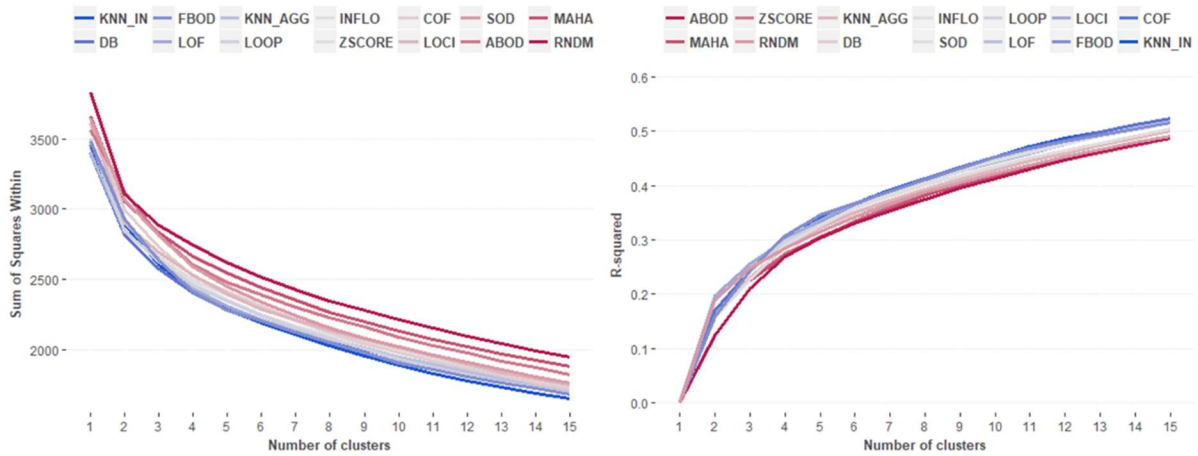


Figure 74: Hierarchical clustering for HeartDisease with 10% of observations removed

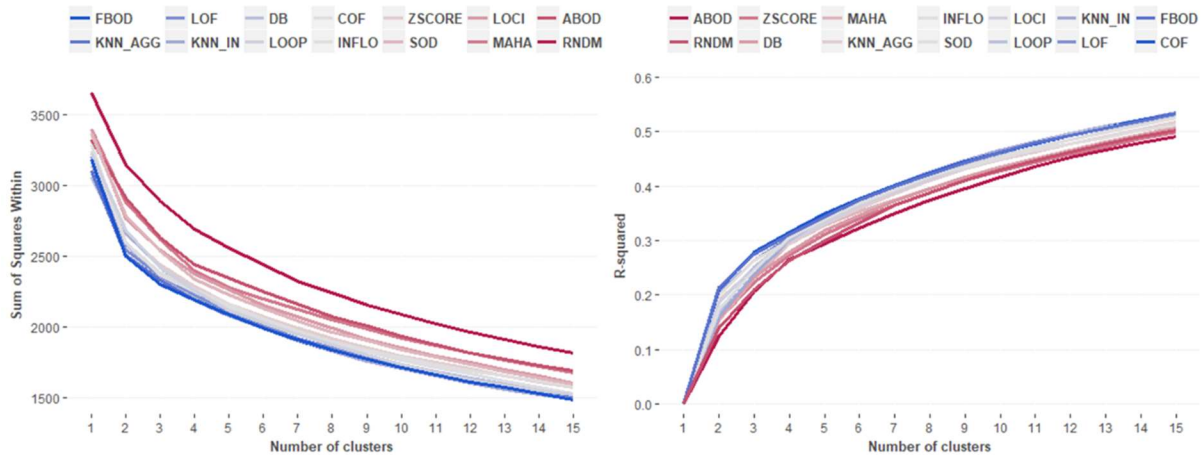


Figure 75: Hierarchical clustering for HeartDisease with 15% of observations removed

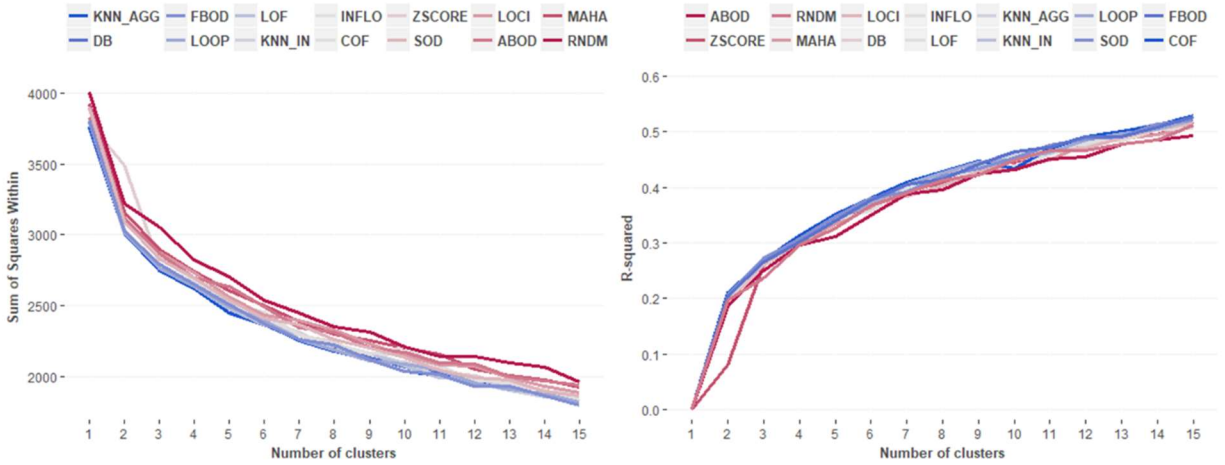


Figure 76: K-means clustering for HeartDisease with 5% of observations removed

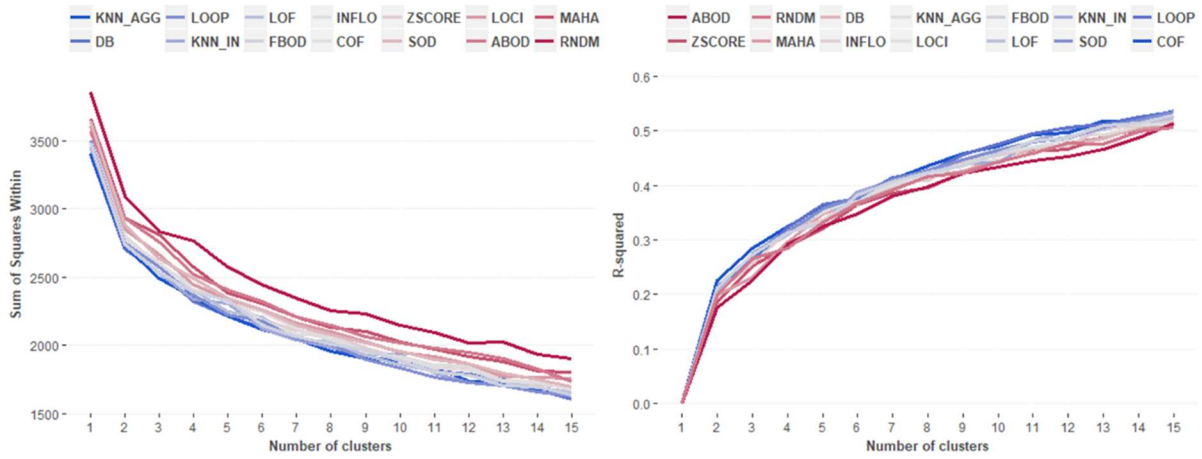


Figure 77: K-means clustering for HeartDisease with 10% of observations removed

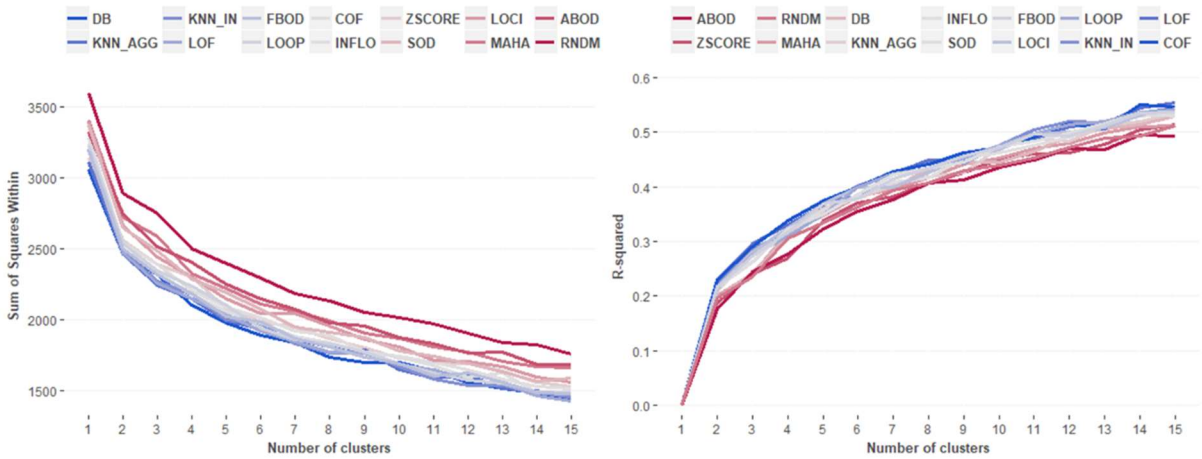


Figure 78: K-means clustering for HeartDisease with 15% of observations removed

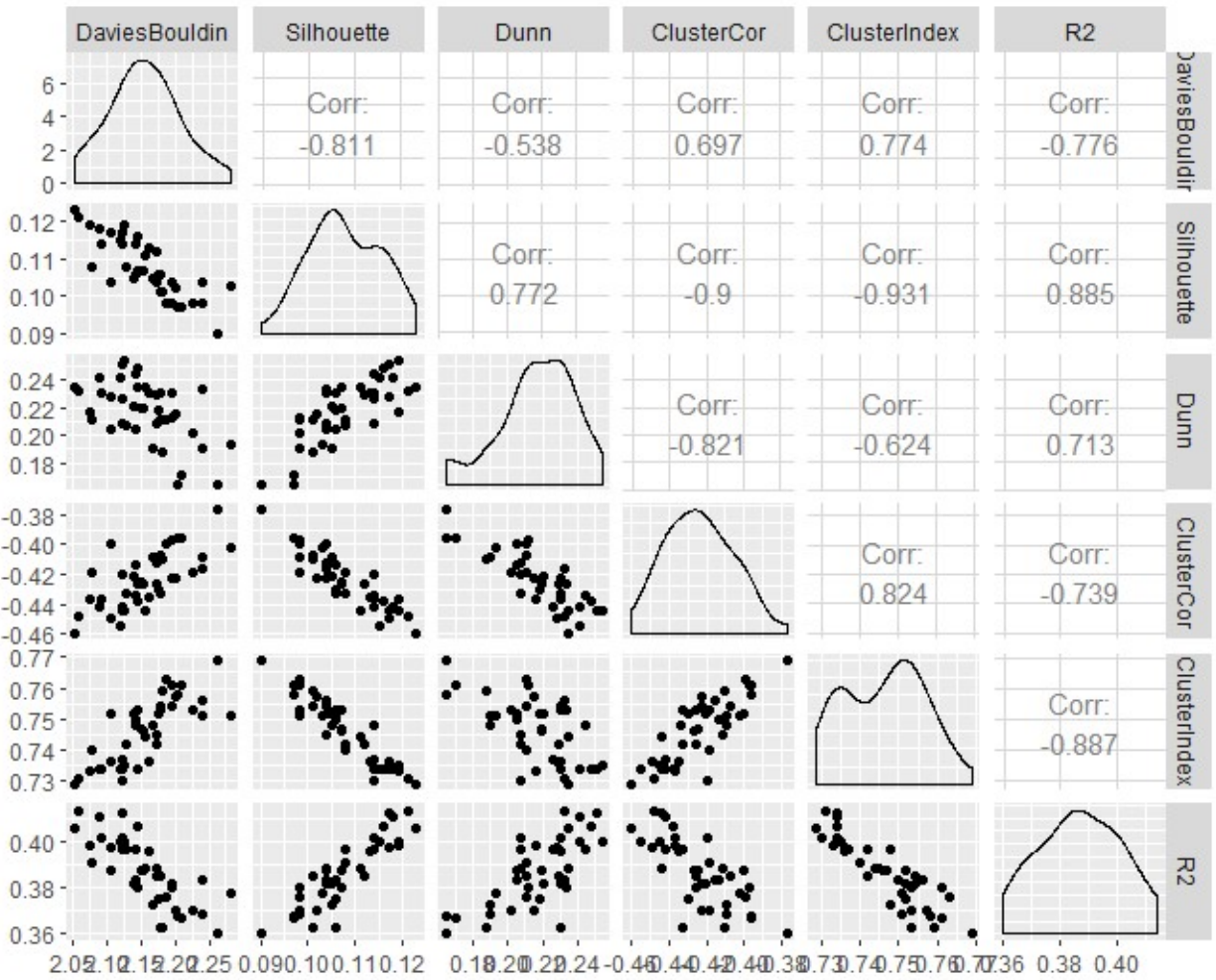


Figure 79: Correlation matrix plot for clustering evaluation metrics for HeartDisease with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
SOD	2.172	0.104	0.208	-0.412	0.745	0.389	2378
LOF	2.152	0.107	0.219	-0.427	0.746	0.388	2320
COF	2.106	0.104	0.205	-0.400	0.752	0.388	2386
LOOP	2.150	0.107	0.219	-0.425	0.747	0.388	2324
FBOD	2.150	0.107	0.219	-0.425	0.747	0.388	2324
KNN_IN	2.179	0.101	0.211	-0.407	0.754	0.385	2342
LOCI	2.141	0.106	0.205	-0.414	0.750	0.384	2402
INFLO	2.138	0.105	0.221	-0.421	0.752	0.382	2361
KNN_AGG	2.195	0.098	0.212	-0.397	0.761	0.380	2335
DB	2.188	0.098	0.211	-0.399	0.763	0.376	2341
ZSCORE	2.227	0.098	0.202	-0.419	0.753	0.370	2387
RNDM	2.204	0.097	0.165	-0.396	0.758	0.368	2556
MAHA	2.209	0.097	0.171	-0.396	0.761	0.367	2481
ABOD	2.262	0.090	0.165	-0.377	0.769	0.360	2448

Table 66: Hierarchical clustering results for HeartDisease with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	2.124	0.119	0.254	-0.445	0.735	0.400	2076
COF	2.073	0.119	0.216	-0.437	0.733	0.399	2171
FBOD	2.104	0.117	0.228	-0.449	0.736	0.398	2096
LOCI	2.129	0.108	0.207	-0.433	0.742	0.397	2177
LOF	2.121	0.114	0.226	-0.442	0.737	0.397	2099
LOOP	2.160	0.113	0.230	-0.435	0.736	0.396	2113
SOD	2.077	0.108	0.211	-0.419	0.740	0.391	2219
INFLO	2.157	0.111	0.235	-0.444	0.744	0.389	2142
DB	2.239	0.104	0.233	-0.416	0.756	0.384	2096
KNN_AGG	2.144	0.107	0.234	-0.426	0.753	0.380	2113
RNDM	2.283	0.103	0.193	-0.402	0.751	0.378	2383
ZSCORE	2.200	0.102	0.215	-0.423	0.757	0.370	2168
MAHA	2.240	0.098	0.190	-0.409	0.751	0.369	2311
ABOD	2.182	0.101	0.188	-0.410	0.759	0.363	2273

Table 67: Hierarchical clustering results for HeartDisease with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	2.057	0.121	0.232	-0.448	0.731	0.414	1926
FBOD	2.123	0.117	0.251	-0.444	0.734	0.413	1869
LOF	2.089	0.118	0.241	-0.442	0.734	0.411	1888
KNN_IN	2.144	0.116	0.248	-0.438	0.734	0.407	1888
LOOP	2.053	0.123	0.235	-0.460	0.729	0.406	1917
LOCI	2.091	0.114	0.230	-0.437	0.734	0.402	2031
SOD	2.121	0.114	0.208	-0.420	0.730	0.402	2015
INFLO	2.119	0.115	0.241	-0.455	0.734	0.400	1945
KNN_AGG	2.141	0.114	0.244	-0.434	0.748	0.397	1873
MAHA	2.172	0.112	0.229	-0.426	0.742	0.385	2094
DB	2.195	0.104	0.230	-0.423	0.754	0.382	1891
ZSCORE	2.175	0.106	0.218	-0.430	0.752	0.375	1962
RNDM	2.167	0.105	0.190	-0.409	0.748	0.373	2295
ABOD	2.177	0.106	0.230	-0.433	0.753	0.363	2117

Table 68: Hierarchical clustering results for HeartDisease with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	2.098	0.123	0.179	-0.435	0.728	0.411	2298
FBOD	2.105	0.119	0.202	-0.452	0.727	0.408	2252
SOD	2.105	0.121	0.187	-0.435	0.725	0.407	2306
LOOP	2.125	0.119	0.181	-0.444	0.731	0.407	2252
KNN_IN	2.151	0.119	0.188	-0.446	0.729	0.407	2259
KNN_AGG	2.143	0.120	0.182	-0.444	0.731	0.406	2239
LOF	2.141	0.119	0.195	-0.444	0.731	0.405	2257
INFLO	2.158	0.115	0.189	-0.439	0.735	0.401	2289
DB	2.138	0.118	0.206	-0.443	0.736	0.400	2251
LOCI	2.132	0.119	0.186	-0.449	0.727	0.400	2340
MAHA	2.147	0.119	0.149	-0.440	0.730	0.396	2368
RNDM	2.152	0.113	0.147	-0.428	0.734	0.393	2432
ZSCORE	2.158	0.118	0.186	-0.437	0.734	0.392	2305
ABOD	2.174	0.114	0.163	-0.433	0.739	0.385	2351

Table 69: K-means clustering results for HeartDisease with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	2.126	0.122	0.176	-0.442	0.720	0.422	2087
LOOP	2.083	0.126	0.214	-0.458	0.722	0.421	2027
SOD	2.091	0.123	0.179	-0.437	0.720	0.415	2130
KNN_IN	2.093	0.122	0.186	-0.463	0.723	0.413	2032
LOF	2.135	0.122	0.204	-0.459	0.724	0.412	2047
FBOD	2.142	0.121	0.208	-0.453	0.728	0.412	2050
LOCI	2.150	0.122	0.193	-0.458	0.724	0.409	2134
KNN_AGG	2.111	0.122	0.181	-0.448	0.733	0.407	2023
INFLO	2.159	0.120	0.211	-0.458	0.729	0.406	2084
DB	2.177	0.121	0.213	-0.451	0.733	0.404	2026
MAHA	2.128	0.120	0.156	-0.446	0.730	0.398	2205
RNDM	2.133	0.118	0.166	-0.436	0.730	0.396	2333
ZSCORE	2.146	0.118	0.200	-0.451	0.734	0.392	2092
ABOD	2.122	0.116	0.168	-0.432	0.743	0.383	2198

Table 70: K-means clustering results for HeartDisease with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	2.078	0.130	0.206	-0.455	0.716	0.432	1867
LOF	2.080	0.131	0.222	-0.461	0.719	0.431	1823
KNN_IN	2.069	0.130	0.226	-0.471	0.715	0.428	1822
LOOP	2.070	0.128	0.189	-0.470	0.715	0.425	1857
LOCI	2.061	0.124	0.207	-0.451	0.722	0.423	1962
FBOD	2.120	0.127	0.205	-0.465	0.719	0.422	1844
SOD	2.044	0.128	0.186	-0.443	0.715	0.420	1956
INFLO	2.060	0.129	0.202	-0.465	0.723	0.417	1892
KNN_AGG	2.084	0.127	0.204	-0.466	0.728	0.414	1820
DB	2.087	0.128	0.210	-0.471	0.729	0.409	1808
MAHA	2.128	0.120	0.186	-0.446	0.727	0.405	2026
RNDM	2.091	0.116	0.191	-0.444	0.733	0.395	2176
ZSCORE	2.126	0.121	0.208	-0.457	0.735	0.392	1909
ABOD	2.151	0.112	0.191	-0.437	0.745	0.384	2046

Table 71: K-means clustering results for HeartDisease with 15% removed observations

9.5.4. Frogs

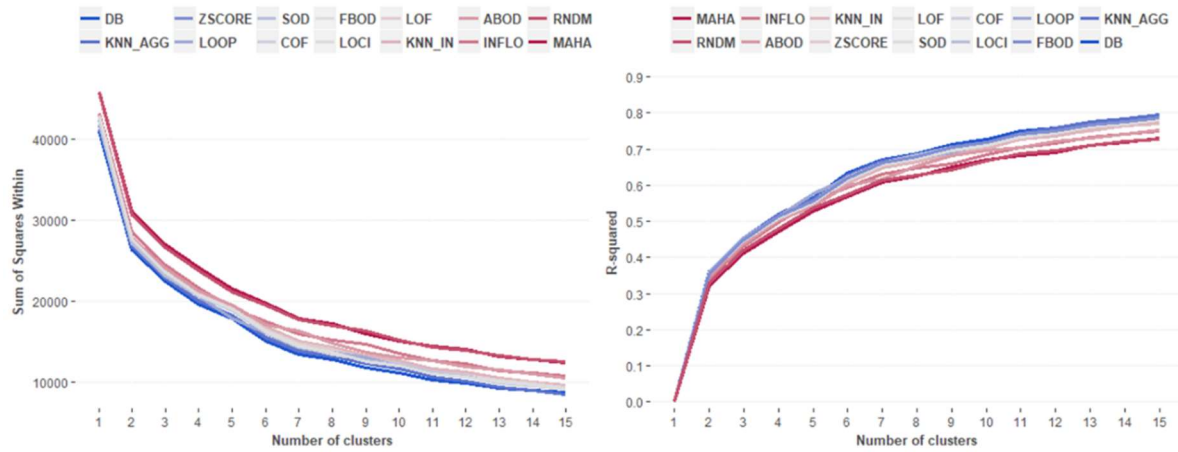


Figure 80: Hierarchical clustering for Frogs with 5% of observations removed

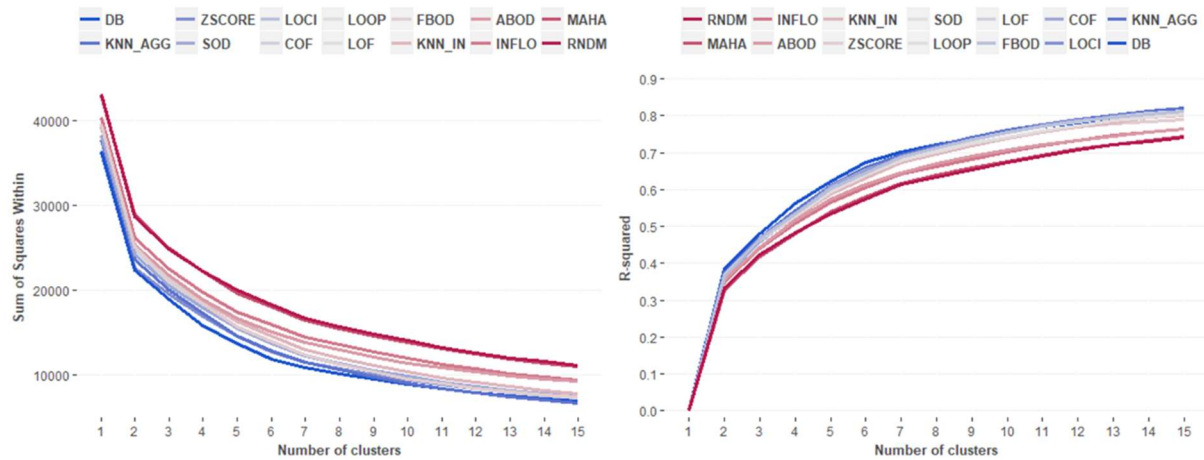


Figure 81: Hierarchical clustering for Frogs with 10% of observations removed

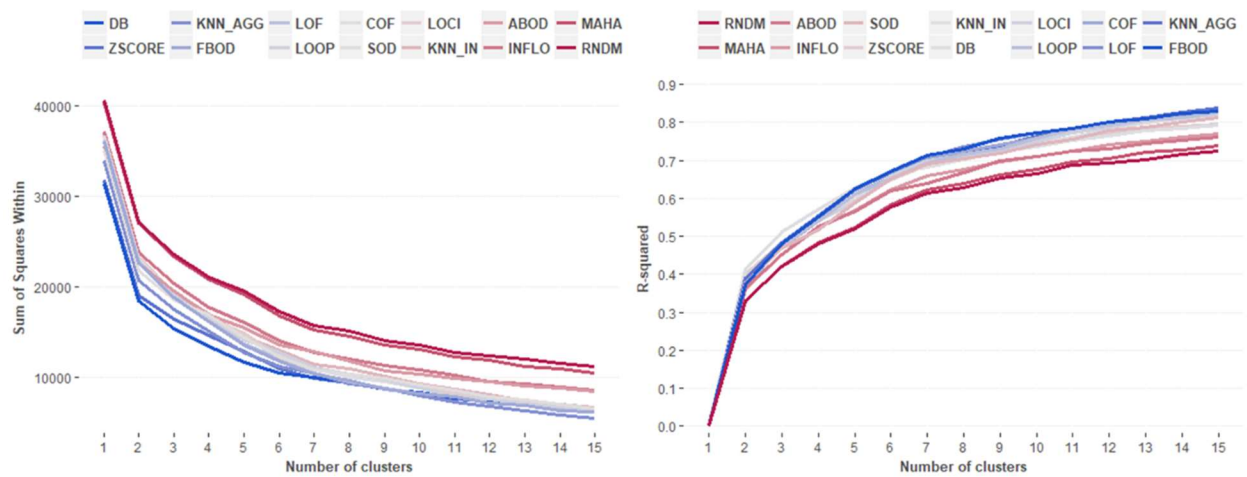


Figure 82: Hierarchical clustering for Frogs with 15% of observations removed

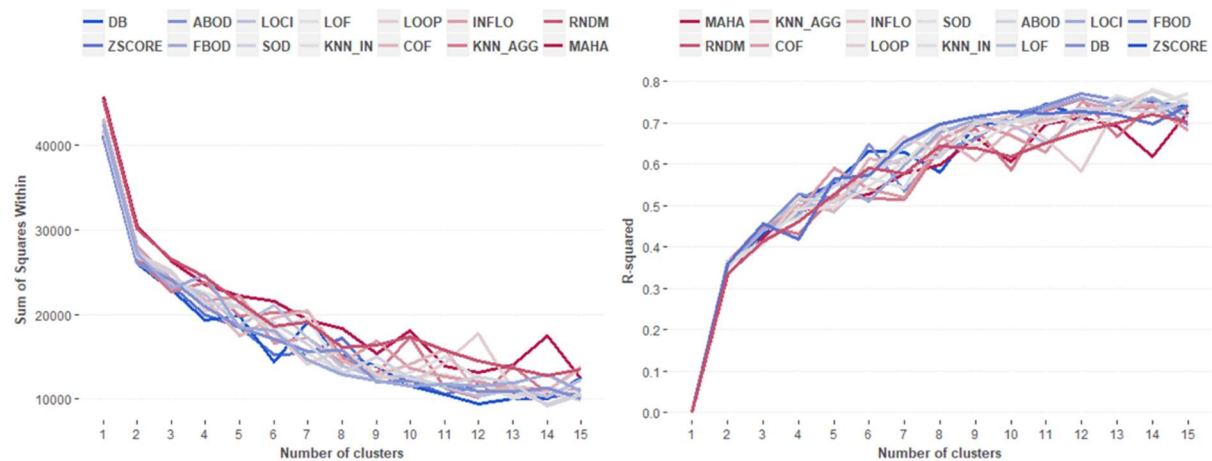


Figure 83: K-means clustering for Frogs with 5% of observations removed

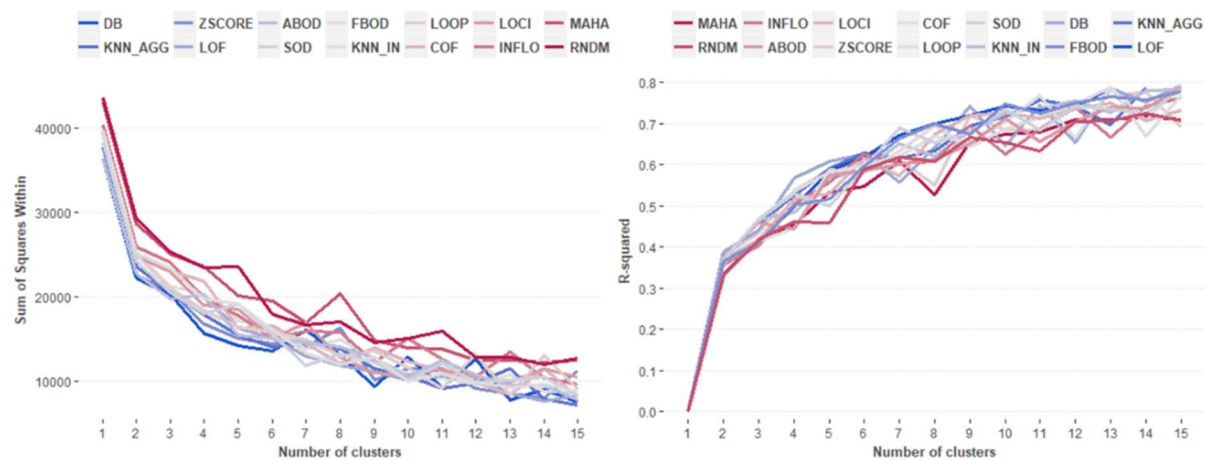


Figure 84: K-means clustering for Frogs with 10% of observations removed

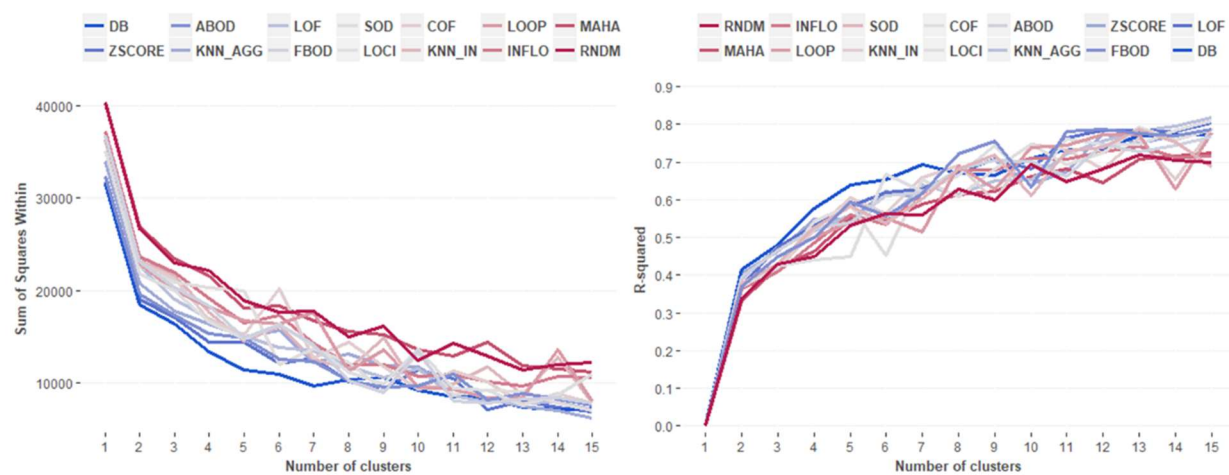


Figure 85: K-means clustering for Frogs with 15% of observations removed

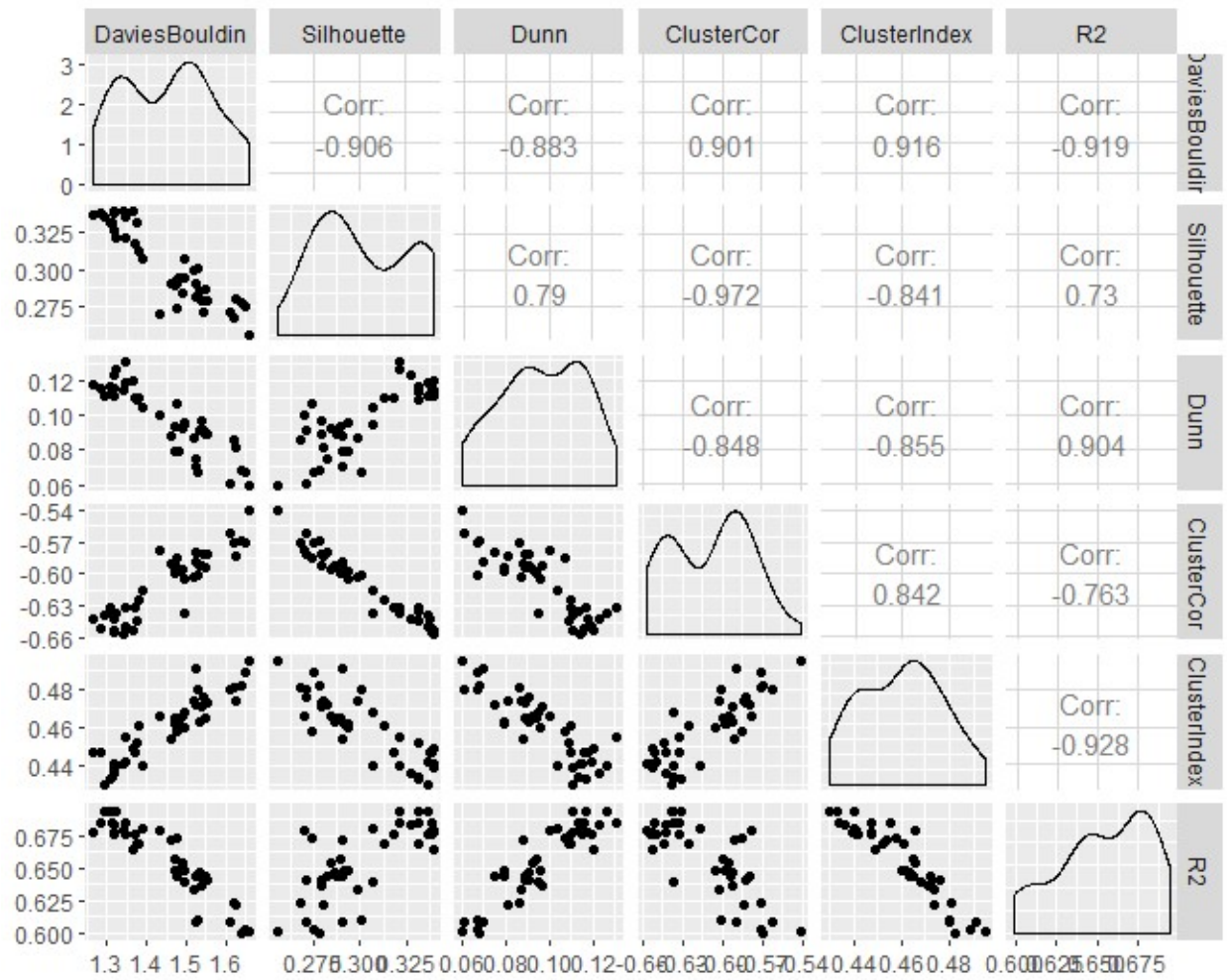


Figure 86: Correlation matrix plot for clustering evaluation metrics for Frogs with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
DB	1.473	0.289	0.094	-0.600	0.465	0.656	14122
KNN_AGG	1.492	0.284	0.092	-0.595	0.466	0.654	14427
FBOD	1.472	0.292	0.079	-0.596	0.462	0.648	15036
LOOP	1.493	0.294	0.096	-0.605	0.460	0.648	14901
LOCI	1.481	0.294	0.079	-0.598	0.462	0.648	15122
COF	1.535	0.286	0.091	-0.593	0.463	0.647	15014
SOD	1.550	0.287	0.089	-0.594	0.465	0.643	14953
LOF	1.474	0.290	0.079	-0.594	0.463	0.643	15315
ZSCORE	1.542	0.271	0.091	-0.582	0.476	0.641	14791
KNN_IN	1.540	0.279	0.097	-0.592	0.471	0.636	15516
ABOD	1.620	0.268	0.086	-0.571	0.481	0.623	16086
INFLO	1.624	0.280	0.081	-0.584	0.474	0.621	16358
RNDM	1.648	0.275	0.067	-0.571	0.489	0.602	18221
MAHA	1.640	0.278	0.068	-0.569	0.482	0.599	18320

Table 72: Hierarchical clustering results for Frogs with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
DB	1.345	0.321	0.131	-0.632	0.455	0.685	11456
KNN_AGG	1.283	0.338	0.116	-0.651	0.447	0.684	11933
LOCI	1.340	0.339	0.114	-0.657	0.441	0.678	12690
COF	1.266	0.337	0.118	-0.642	0.447	0.677	12729
FBOD	1.319	0.339	0.111	-0.653	0.439	0.677	12771
LOF	1.317	0.338	0.115	-0.654	0.441	0.676	12765
LOOP	1.345	0.336	0.119	-0.649	0.442	0.676	12761
SOD	1.374	0.331	0.109	-0.644	0.452	0.669	12677
ZSCORE	1.379	0.313	0.110	-0.625	0.461	0.668	12069
KNN_IN	1.365	0.339	0.120	-0.653	0.449	0.664	13322
ABOD	1.496	0.307	0.095	-0.637	0.468	0.639	14165
INFLO	1.517	0.299	0.087	-0.603	0.474	0.634	14731
MAHA	1.528	0.301	0.067	-0.601	0.480	0.610	16753
RNDM	1.526	0.290	0.070	-0.589	0.491	0.608	16873

Table 73: Hierarchical clustering results for Frogs with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
FBOD	1.293	0.336	0.111	-0.638	0.430	0.694	11045
KNN_AGG	1.321	0.321	0.126	-0.637	0.440	0.693	10428
LOF	1.305	0.331	0.117	-0.631	0.433	0.693	11121
COF	1.314	0.331	0.113	-0.636	0.434	0.685	11561
LOOP	1.318	0.327	0.123	-0.643	0.436	0.683	11526
LOCI	1.389	0.307	0.104	-0.616	0.440	0.680	11585
DB	1.432	0.270	0.100	-0.578	0.466	0.679	10118
KNN_IN	1.368	0.318	0.110	-0.632	0.447	0.676	11956
ZSCORE	1.477	0.274	0.107	-0.585	0.458	0.673	10396
SOD	1.461	0.291	0.088	-0.591	0.454	0.671	11568
INFLO	1.524	0.282	0.075	-0.579	0.472	0.644	13265
ABOD	1.551	0.279	0.089	-0.582	0.473	0.640	12848
MAHA	1.611	0.271	0.061	-0.562	0.480	0.609	15782
RNDM	1.658	0.256	0.060	-0.541	0.495	0.601	16227

Table 74: Hierarchical clustering results for Frogs with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
ZSCORE	1.592	0.270	0.061	-0.531	0.509	0.628	15362
FBOD	1.525	0.268	0.061	-0.511	0.527	0.627	15859
DB	1.559	0.268	0.074	-0.510	0.538	0.626	15355
LOCI	1.556	0.261	0.050	-0.509	0.518	0.624	16131
LOF	1.516	0.281	0.059	-0.517	0.523	0.622	16210
ABOD	1.612	0.243	0.068	-0.494	0.536	0.621	15501
KNN_IN	1.528	0.260	0.064	-0.499	0.532	0.618	16298
SOD	1.562	0.241	0.062	-0.457	0.564	0.615	16146
LOOP	1.550	0.254	0.063	-0.478	0.558	0.613	16392
INFLO	1.599	0.268	0.059	-0.516	0.532	0.610	16823
COF	1.595	0.262	0.067	-0.495	0.547	0.607	16747
KNN_AGG	1.591	0.253	0.068	-0.470	0.578	0.596	16856
RNDM	1.672	0.261	0.041	-0.512	0.527	0.590	18597
MAHA	1.700	0.260	0.040	-0.518	0.524	0.584	19035

Table 75: K-means clustering results for Frogs with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOF	1.458	0.282	0.071	-0.527	0.490	0.654	13634
KNN_AGG	1.477	0.286	0.082	-0.536	0.506	0.645	13388
FBOD	1.479	0.258	0.066	-0.479	0.526	0.639	14304
DB	1.547	0.259	0.076	-0.496	0.525	0.639	13130
KNN_IN	1.552	0.276	0.064	-0.543	0.500	0.638	14375
SOD	1.531	0.283	0.071	-0.526	0.515	0.634	13987
LOOP	1.534	0.274	0.068	-0.522	0.512	0.632	14476
COF	1.515	0.253	0.064	-0.485	0.533	0.629	14630
ZSCORE	1.613	0.251	0.074	-0.489	0.533	0.627	13572
LOCI	1.555	0.250	0.064	-0.479	0.537	0.627	14690
ABOD	1.611	0.245	0.066	-0.494	0.537	0.617	13942
INFLO	1.602	0.258	0.050	-0.506	0.525	0.613	15606
RNDM	1.674	0.252	0.042	-0.489	0.537	0.591	17799
MAHA	1.716	0.245	0.040	-0.489	0.533	0.591	17599

Table 76: K-means clustering results for Frogs with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
DB	1.552	0.262	0.084	-0.517	0.501	0.663	10606
LOF	1.519	0.273	0.074	-0.519	0.495	0.656	12442
FBOD	1.441	0.283	0.074	-0.514	0.500	0.650	12858
ZSCORE	1.571	0.245	0.079	-0.485	0.515	0.640	11434
KNN_AGG	1.516	0.272	0.075	-0.485	0.524	0.636	12354
ABOD	1.594	0.260	0.077	-0.534	0.507	0.635	11779
LOCI	1.460	0.255	0.072	-0.451	0.547	0.634	13255
COF	1.481	0.270	0.072	-0.476	0.534	0.630	13601
KNN_IN	1.480	0.270	0.067	-0.512	0.536	0.628	13725
SOD	1.552	0.256	0.066	-0.481	0.535	0.626	13134
LOOP	1.537	0.253	0.075	-0.459	0.550	0.620	13829
INFLO	1.635	0.263	0.050	-0.512	0.524	0.617	14241
MAHA	1.629	0.262	0.045	-0.490	0.545	0.591	16517
RNDM	1.667	0.247	0.054	-0.501	0.534	0.589	16594

Table 77: K-means clustering results for Frogs with 15% removed observations

9.5.5. WineQuality

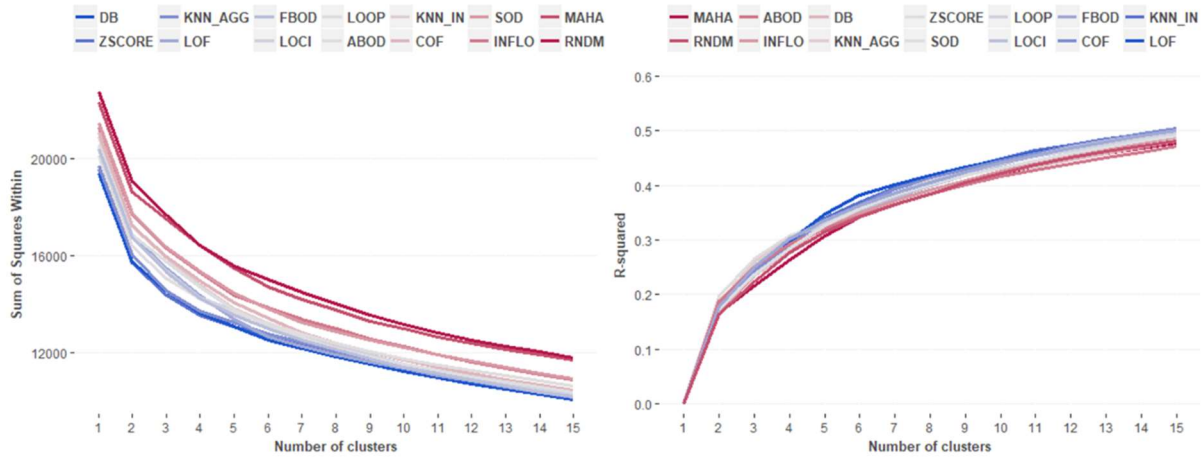


Figure 87: Hierarchical clustering for WineQuality with 5% removed observations

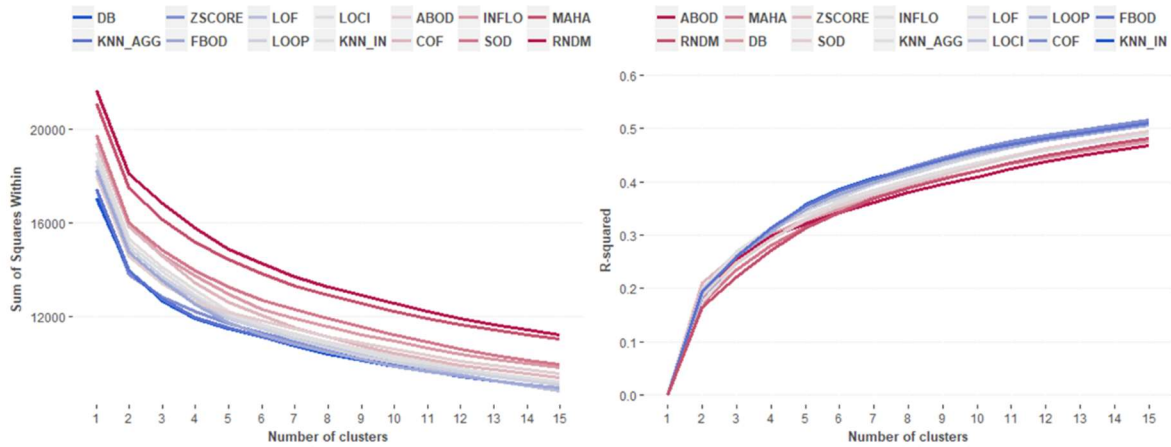


Figure 88: Hierarchical clustering for WineQuality with 10% removed observations

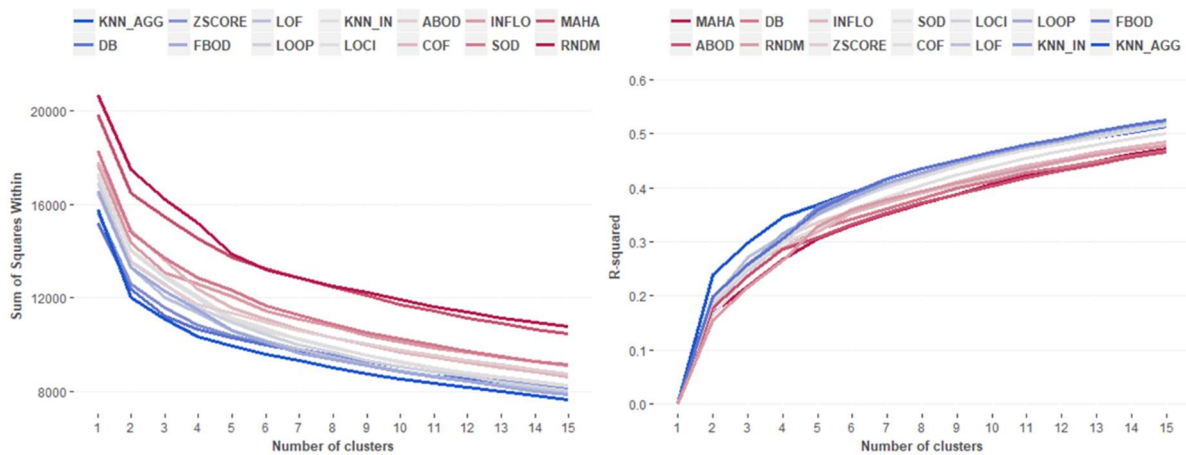


Figure 89: Hierarchical clustering for WineQuality with 15% removed observations

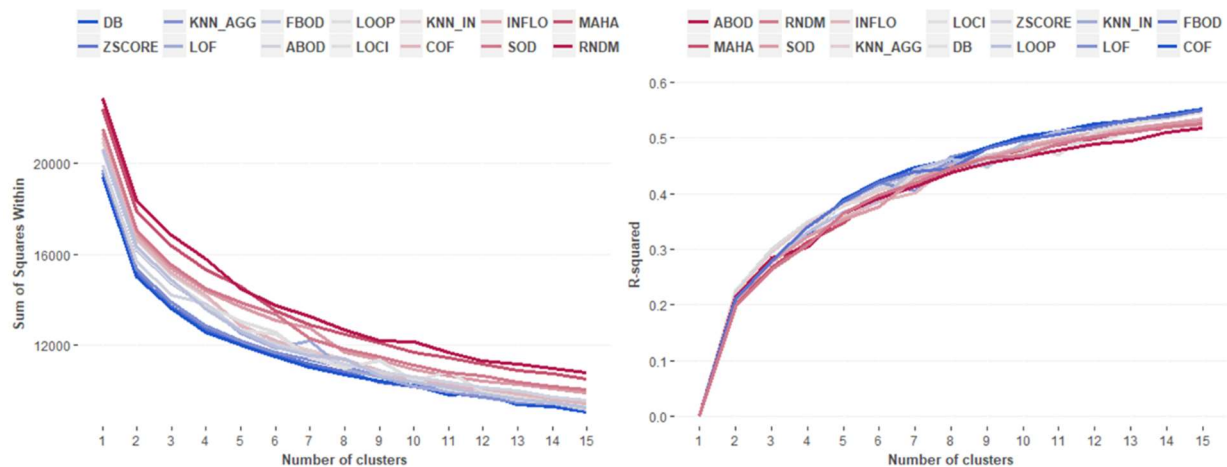


Figure 90: K-means clustering for WineQuality with 5% removed observations

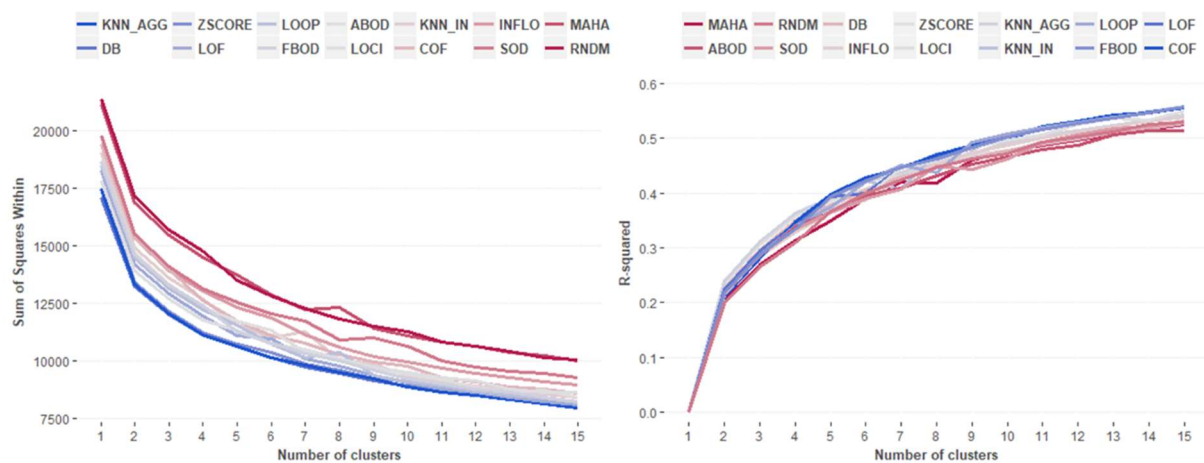


Figure 91: K-means clustering for WineQuality with 10% removed observations

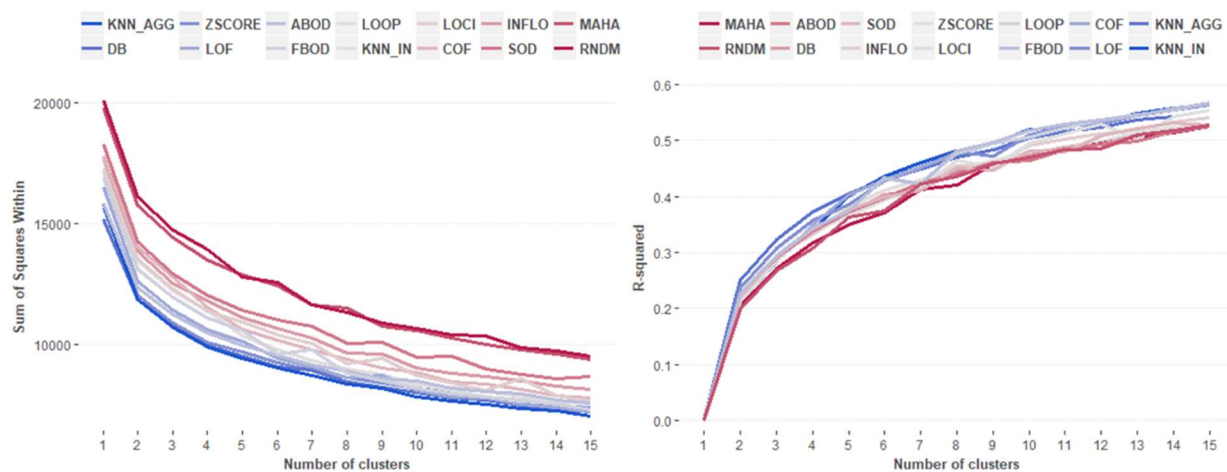


Figure 92: K-means clustering for WineQuality with 15% removed observations

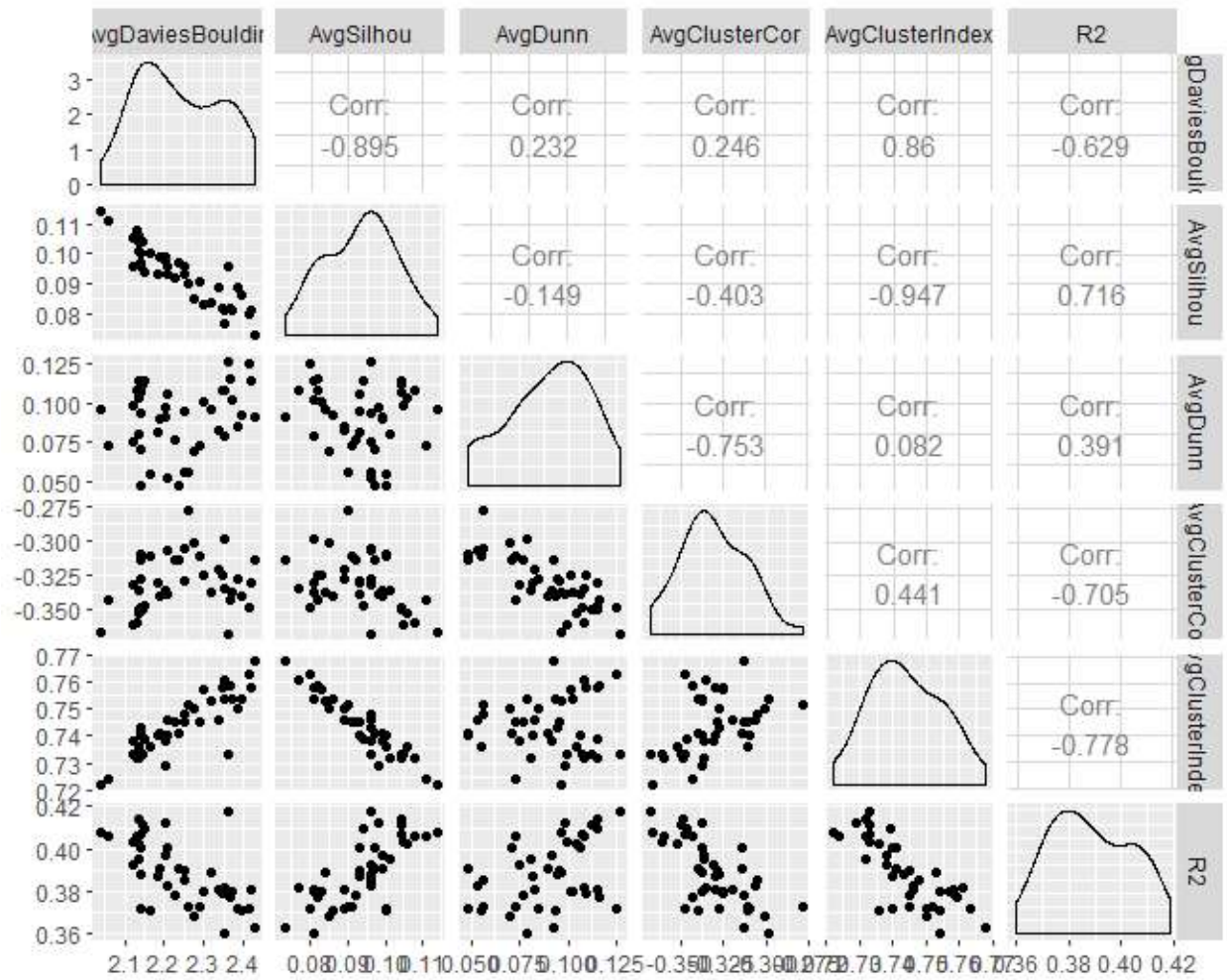


Figure 93: Correlation matrix plot for clustering evaluation metrics for WineQuality with hierarchical clustering

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
LOF	2.201	0.099	0.091	-0.336	0.738	0.397	12353
KNN_IN	2.134	0.101	0.081	-0.336	0.732	0.395	12656
COF	2.121	0.096	0.075	-0.331	0.738	0.393	12795
FBOD	2.188	0.099	0.090	-0.340	0.741	0.391	12416
LOCI	2.252	0.093	0.095	-0.329	0.745	0.390	12513
LOOP	2.139	0.096	0.094	-0.327	0.743	0.388	12600
SOD	2.207	0.096	0.053	-0.306	0.746	0.383	13249
ZSCORE	2.369	0.081	0.102	-0.337	0.754	0.380	12131
KNN_AGG	2.297	0.083	0.101	-0.325	0.757	0.380	12200
DB	2.349	0.082	0.109	-0.325	0.758	0.379	12052
INFLO	2.227	0.092	0.077	-0.313	0.745	0.378	13252
ABOD	2.386	0.089	0.085	-0.327	0.750	0.372	12628
RNDM	2.139	0.100	0.048	-0.310	0.740	0.372	14311
MAHA	2.273	0.085	0.070	-0.301	0.750	0.368	14121

Table 78: Hierarchical clustering results for WineQuality with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	2.042	0.114	0.096	-0.367	0.722	0.408	11265
FBOD	2.129	0.108	0.108	-0.360	0.732	0.406	10845
COF	2.060	0.111	0.073	-0.343	0.724	0.406	11519
LOOP	2.123	0.105	0.099	-0.361	0.733	0.403	11029
LOCI	2.135	0.106	0.104	-0.352	0.736	0.402	11152
LOF	2.206	0.093	0.106	-0.338	0.740	0.401	10929
KNN_AGG	2.319	0.084	0.096	-0.337	0.753	0.389	10669
INFLO	2.184	0.093	0.082	-0.330	0.740	0.387	11893
SOD	2.250	0.096	0.056	-0.305	0.748	0.385	12141
ZSCORE	2.420	0.081	0.115	-0.330	0.758	0.381	10813
DB	2.366	0.082	0.116	-0.343	0.759	0.377	10637
MAHA	2.291	0.091	0.073	-0.311	0.745	0.373	13248
RNDM	2.166	0.100	0.055	-0.311	0.736	0.371	13637
ABOD	2.394	0.086	0.093	-0.340	0.754	0.371	11330

Table 79: Hierarchical clustering results for WineQuality with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_AGG	2.362	0.096	0.127	-0.368	0.733	0.418	9193
FBOD	2.138	0.104	0.115	-0.349	0.732	0.414	9710
KNN_IN	2.201	0.098	0.098	-0.338	0.729	0.412	10182
LOOP	2.144	0.104	0.112	-0.350	0.733	0.411	9960
LOF	2.150	0.094	0.115	-0.347	0.739	0.410	9732
LOCI	2.140	0.104	0.107	-0.348	0.733	0.407	10197
COF	2.142	0.097	0.071	-0.314	0.741	0.401	10669
SOD	2.237	0.097	0.048	-0.314	0.741	0.391	11139
ZSCORE	2.352	0.077	0.108	-0.335	0.761	0.382	9683
INFLO	2.338	0.089	0.083	-0.320	0.746	0.381	10951
RNDM	2.262	0.090	0.056	-0.278	0.752	0.373	12962
DB	2.413	0.080	0.125	-0.348	0.763	0.372	9544
ABOD	2.429	0.073	0.092	-0.313	0.768	0.363	10475
MAHA	2.350	0.081	0.079	-0.298	0.754	0.360	12664

Table 80: Hierarchical clustering results for WineQuality with 15% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.981	0.132	0.041	-0.350	0.703	0.442	11778
FBOD	1.957	0.133	0.040	-0.364	0.705	0.439	11557
LOF	1.967	0.134	0.041	-0.360	0.707	0.438	11509
KNN_IN	1.971	0.132	0.044	-0.355	0.707	0.438	11768
LOOP	2.021	0.129	0.049	-0.355	0.708	0.434	11653
ZSCORE	2.072	0.123	0.057	-0.366	0.718	0.432	11123
DB	2.083	0.122	0.064	-0.367	0.721	0.430	11058
LOCI	2.012	0.128	0.042	-0.357	0.711	0.430	11689
KNN_AGG	2.068	0.123	0.051	-0.365	0.719	0.428	11257
INFLO	2.033	0.125	0.046	-0.341	0.712	0.424	12258
SOD	2.054	0.125	0.025	-0.331	0.715	0.423	12373
RNDM	2.013	0.127	0.025	-0.320	0.712	0.420	13249
MAHA	2.010	0.125	0.041	-0.340	0.711	0.419	12978
ABOD	2.104	0.121	0.043	-0.354	0.722	0.416	11611

Table 81: K-means clustering results for WineQuality with 5% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
COF	1.941	0.134	0.041	-0.358	0.699	0.447	10722
LOF	1.975	0.132	0.065	-0.372	0.705	0.446	10104
FBOD	1.954	0.133	0.055	-0.365	0.704	0.445	10317
LOOP	1.988	0.134	0.057	-0.374	0.702	0.445	10259
KNN_IN	1.998	0.132	0.053	-0.361	0.703	0.444	10568
KNN_AGG	2.051	0.126	0.054	-0.380	0.712	0.443	9724
LOCI	2.010	0.129	0.051	-0.364	0.709	0.439	10474
ZSCORE	2.065	0.124	0.063	-0.376	0.717	0.435	9864
INFLO	2.020	0.127	0.041	-0.352	0.708	0.430	11053
DB	2.092	0.122	0.070	-0.383	0.720	0.430	9739
SOD	2.080	0.125	0.027	-0.331	0.716	0.423	11390
RNDM	2.022	0.127	0.037	-0.333	0.711	0.422	12340
ABOD	2.066	0.123	0.048	-0.360	0.724	0.419	10341
MAHA	2.051	0.123	0.035	-0.336	0.713	0.417	12317

Table 82: K-means clustering results for WineQuality with 10% removed observations

	Avg Davies-Bouldin	Avg Silhou	Avg Dunn	Avg ClusterCor	Avg ClusterIndex	Avg R ²	Avg SSW
KNN_IN	1.951	0.137	0.051	-0.377	0.695	0.456	9414
KNN_AGG	2.054	0.127	0.076	-0.392	0.705	0.454	8613
LOF	1.988	0.131	0.062	-0.379	0.702	0.453	9019
COF	1.926	0.137	0.041	-0.361	0.697	0.452	9759
FBOD	2.002	0.132	0.054	-0.369	0.699	0.452	9268
LOOP	1.969	0.136	0.056	-0.373	0.700	0.450	9303
LOCI	2.025	0.129	0.052	-0.358	0.712	0.433	9737
ZSCORE	2.101	0.122	0.066	-0.382	0.719	0.433	8875
INFLO	2.033	0.129	0.050	-0.359	0.708	0.432	10059
SOD	2.024	0.128	0.029	-0.338	0.713	0.428	10463
DB	2.077	0.124	0.069	-0.391	0.723	0.425	8727
ABOD	2.091	0.122	0.044	-0.376	0.721	0.424	9143
RNDM	2.040	0.125	0.030	-0.325	0.713	0.416	11745
MAHA	2.047	0.122	0.041	-0.332	0.720	0.414	11602

Table 83: K-means clustering results for WineQuality with 15% removed observations