

The copyright © of this thesis belongs to its rightful author and/or other copyright owner. Copies can be accessed and downloaded for non-commercial or learning purposes without any charge and permission. The thesis cannot be reproduced or quoted as a whole without the permission from its rightful owner. No alteration or changes in format is allowed without permission from its rightful owner.



**STATISTICAL PERFORMANCE OF AGGLOMERATIVE
HIERARCHICAL CLUSTERING TECHNIQUE VIA PAIRING
OF CORRELATION-BASED DISTANCES AND LINKAGE
METHODS**

NURSHAZIANA BINTI MOHAMAD SHAMSURI



MASTER OF SCIENCE (STATISTICS)

UNIVERSITI UTARA MALAYSIA

2025



Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI
(Certification of thesis / dissertation)

Kami, yang bertandatangan, memperakukan bahawa
(We, the undersigned, certify that)

NURSHAZIANA MOHAMAD SHAMSURI

calon untuk Ijazah

(candidate for the degree of)

MASTER OF SCIENCE (STATISTICS)

telah mengemukakan tesis / disertasi yang bertajuk:

(has presented his/her thesis / dissertation of the following title):

**“STATISTICAL PERFORMANCE OF AGGLOMERATIVE HIERARCHICAL CLUSTERING
TECHNIQUE VIA PAIRING OF CORRELATION-BASED DISTANCES AND LINKAGE METHODS”**

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
(as it appears on the title page and front cover of the thesis / dissertation).

Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **05 Mac 2024.**

That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:
05 March 2024.

Pengerusi Viva:
(Chairman for VIVA)

Assoc. Prof. Dr. Suhaida Abdullah

Tandatangan
(Signature)

Pemeriksa Luar:
(External Examiner)

**Assoc. Prof. Ts Dr. Ahmad Zia Ul-Saufie
Mohamad Japeri**

Tandatangan
(Signature)

Pemeriksa Dalam:
(Internal Examiner)

Dr. Malina Zulkifli

Tandatangan
(Signature)

Nama Penyelia/Penyelia-penyelia:
(Name of Supervisor/Supervisors)

Assoc. Prof. Dr. Shamshuritawati Sharif

Tandatangan
(Signature)

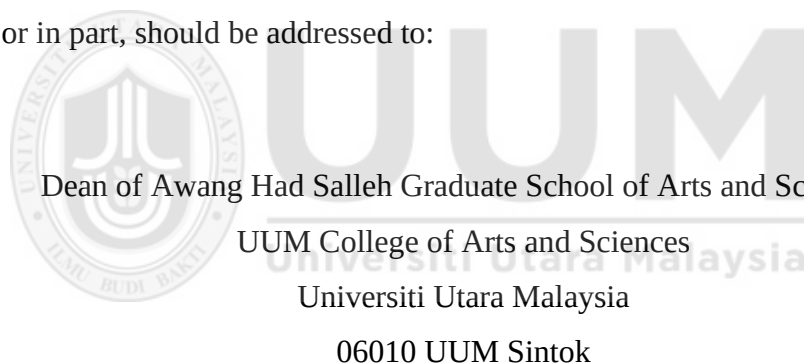
Tarikh:

(Date) **05 March 2024**

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:



Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Teknik pengelompokan berhierarki aglomeratif membantu mengelompokkan objek yang tidak diketahui ke dalam beberapa kluster. Kualiti pengelompokan dapat dicapai apabila kluster tersebut homogen secara dalaman dan heterogen secara luaran. Walau bagaimanapun, teknik ini sensitif terhadap pemilihan ukuran jarak dan kaedah penghubung, terutamanya jarak yang berasaskan korelasi kerana ia dipengaruhi oleh data terpencil. Pemilihan pasangan yang salah boleh mendorong kepada pembentukan kluster yang lemah dan prestasi pengelompokan yang rendah. Oleh itu, objektif utama adalah untuk mengukur prestasi berstatistik pengelompokan hierarki aglomeratif menggunakan empat jarak berasaskan korelasi dan empat kaedah penghubung, yang diuji melalui kajian simulasi dan aplikasi data sebenar. Empat jarak berasaskan korelasi tersebut; *Pearson*, *Spearman*, *Kendall's Tau* dan *Winsorized* manakala empat kaedah penghubung; *Average*, *Ward's*, *Complete* dan *Single*. Kajian simulasi dijalankan dengan mempertimbangkan pelbagai kondisi data; bilangan pemboleh ubah, saiz sampel, peratus data pencilan, dan taburan data untuk mengukur prestasi dengan menggunakan Pekali Korelasi Kofenetik (CCC). Untuk mengesahkan model pengelompokan tersebut terhadap data sebenar, algoritma *Spearman-average* digunakan untuk mengelompokkan data lembangan Sungai Juru berdasarkan lima parameter kualiti air. Kajian simulasi menunjukkan bahawa *Spearman-average* berprestasi baik di bawah taburan normal dan data tercemar manakala *Winsorized-average* di bawah taburan Gamma. Lima jadual ringkasan untuk memilih algoritma pengelompokan yang sesuai mengikut taburan data telah dihasilkan. Pengesahan data sebenar menghasilkan lima kluster dengan CCC ialah 0.76 dan indeks Calinski dan Harabasz ialah 26.879. Sebagai tambahan, model pengelompokan juga mampu mengenal pasti masa apabila tahap pencemaran sungai tertinggi dan terendah berlaku. Kesimpulan, prestasi pengelompokan bergantung kepada kondisi data. Pengelompokan hierarki aglomeratif sesuai dilaksanakan apabila bilangan pembolehubah tidak melebihi 20 dan saiz sampel kurang daripada 500.

Kata Kunci: Pengelompokan aglomeratif hierarki, Jarak berasaskan korelasi, Kaedah penghubung, Pekali Korelasi Kofenetik

Abstract

Agglomerative hierarchical clustering technique assists to group unknown objects into several clusters. The quality of clustering can be achieved when the clusters are internally homogeneous and externally heterogeneous. However, the technique is sensitive to the choice of the distance measure and linkage method, particularly to the correlation-based distance because it is influenced by outliers. Choosing the wrong pairing may lead to the poor cluster formation and low clustering performance. Therefore, the main objective was to measure the statistical performance of agglomerative hierarchical clustering using four correlation-based distances and four linkage methods, tested through simulation study and real data application. The four correlation-based distances; Pearson, Spearman, Kendall's Tau and Winsorized while the four linkage methods; Average, Ward's, Complete and Single. The simulation study was conducted under various data conditions; number of variables, sample sizes, percentages of outliers, and data distribution to measure the performance using Cophenetic Correlation Coefficient (CCC). To validate the clustering model on real data, the *Spearman-average* algorithm was applied to cluster Juru river basin data based on five water quality parameters. The simulation study showed that the *Spearman-average* performed well under normal distribution and contaminated data, while the *Winsorized-average* performed well under Gamma distribution. Five tables of summary for choosing appropriate clustering algorithms according to data distribution were produced. The real data validation produced five clusters, with the CCC of 0.76 and the Calinski and Harabasz index of 26.879. In addition, the clustering model was also able to identify the time when the highest and lowest level of river contamination occurred. As a conclusion, the performance of clustering depends on the conditions of the data. Agglomerative hierarchical clustering is suitable to be performed when the number of variables is not more than 20 and the sample size smaller than 500.

Keywords: Agglomerative hierarchical clustering, Correlation-based distances, Linkage methods, Cophenetic Correlation Coefficient.

Acknowledgement

Firstly, I would like to say Alhamdulillah, all praise and thanks to Allah SWT, the Almighty, who has given me the opportunity and strength to complete this thesis. Without Him and His blessings, I may not have been able to finish this journey, which has given me such meaningful experiences and has helped me improved myself to be better in many aspects.

Next, I want to express my deep appreciation and heartfelt gratitude to all those who contributed to the completion of this thesis. Studying under the guidance of Associate Professor Dr. Shamshuritawati Binti Sharif has been a pleasure, and words cannot adequately convey my gratitude for her invaluable support in helping me reach my academic goal. Without her unwavering assistance, my thesis would not have come to fruition.

My heartfelt gratitude also goes out to my parents, and all my relatives for their unwavering love and support. Without their presence and encouragement in my life, achieving this goal would have been impossible. Most importantly, I dedicate this work to my parents.

My time at Universiti Utara Malaysia (UUM) has been exceptionally enjoyable, not only due to its stunning natural surroundings but also thanks to the helpfulness of the university staff.

Lastly, I want to express my gratitude to all my friends for their constant encouragement throughout my academic journey.

Table of Contents

Permission to Use.....	ii
Abstrak	iii
Abstract	iv
Acknowledgement.....	v
Table of Contents	vi
List of Tables.....	ix
List of Figures	xii
List of Abbreviations.....	xiii
 CHAPTER ONE INTRODUCTION	 1
1.1 Introduction.....	1
1.2 Issue on Agglomerative Hierarchical Clustering	6
1.3 Research Problem Statement	10
1.4 Research Questions	13
1.5 Research Objectives	13
1.6 Research Significant.....	14
1.7 Research Scope	15
 CHAPTER TWO LITERATURE REVIEW	 16
2.1 Introduction	16
2.2 Hierarchical clustering	16
2.3 Distance measure	22
2.4 Parametric Method : Pearson Correlation Coefficient (PCC).....	25
2.5 Outliers.....	27
2.6 Non-parametric methods.....	31
2.6.1 Spearman Correlation Coefficient	31
2.6.2 Kendall's Tau correlation	33
2.7 Winsorized Correlation Robust Estimator	33
2.8 Previous studies on the comparison of correlation coefficients.....	35
2.9 Linkage Method	37
2.9.1 Average Linkage.....	37
2.9.2 Ward's Method	38
2.9.3 Complete Linkage.....	38

2.9.4 Single Linkage	39
2.9.5 Past studies on linkage methods	40
3.0 Water Quality Assessment using Cluster Analysis.....	44
3.1 Clustering validation measures	47
CHAPTER THREE RESEARCH METHODOLOGY	52
3.1 Introduction	52
3.2 Procedure implemented.....	52
3.3 Simulation study	58
3.4 Correlation-based distances	62
3.4.1 Pearson Correlation Coefficient (PCC)	62
3.4.2 Spearman correlation	63
3.4.3 Kendall's Tau correlation	64
3.4.4 Winsorized Correlation.....	65
3.5 Linkage method.....	66
3.5.1 Average linkage	66
3.5.2 Ward's Method	67
3.5.3 Complete Linkage.....	68
3.5.4 Single Linkage	69
3.6 Cophenetic Correlation Coefficient (CCC).....	69
3.7 Real Data Application on the sample of Juru River Basin Water Quality.....	70
3.8 Calinski & Harabasz (CH) index	73
CHAPTER FOUR ANALYSIS AND FINDINGS	74
4.1 Introduction	74
4.2 Results of simulation study	74
4.2.1 Clustering Algorithm Under Normal Distribution	76
4.2.2 Clustering Algorithm under Mild and Extreme Contaminated Distribution	83
4.2.3 Clustering Algorithm under Non Normal Gamma Distribution.....	113
4.3 Real Data Screening	126
4.4 Real data analysis using Juru River Basin water quality parameter	128
4.4.1 Descriptive analysis of five water quality parameter.....	129
4.4.2 Correlation Matrix	132
4.4.3 Clustering Model Analysis	133

CHAPTER FIVE CONCLUSION	145
5.1 Conclusions	145
5.2 Significance of Study	148
5.3 Limitation and Constraints faced in Study.....	148
5.4 Recommendation for Future Researchers	148
REFERENCES.....	152



List of Tables

Table 2.1	Past studies on the issue on outliers in clustering.....	29
Table 2.2	Summary on the comparison of correlation coefficients.....	35
Table 2.3	Comparison of linkage methods	39
Table 2.4	Summary of linkage method literatures.....	42
Table 3.1	Full detailed process of the research procedures	54
Table 3.2	List of general commands for simulated data in MATLAB	60
Table 4.1	Rule of Thumb for the interpretation of correlation coefficient	75
Table 4.2	Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 3$	76
Table 4.3	Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 5$	78
Table 4.4	Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 10$	79
Table 4.5	Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 20$	80
Table 4.6	Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 30$	82
Table 4.7	Summary of appropriate clustering algorithms under normal distribution	83
Table 4.8	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu = 3$) contaminated observations when the number of variables, $p = 3$	84
Table 4.9	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu = 3$) contaminated observations when the number of variables, $p = 5$	87
Table 4.10	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu = 3$) contaminated observations when the number of variables, $p = 10$	90
Table 4.11	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu = 3$) contaminated observations when the number of variables, $p = 20$	93

Table 4.12	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu = 3$) contaminated observations when the number of variables, $p = 30$	96
Table 4.13	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu = 5$) contaminated observations when the number of variables, $p = 3$	98
Table 4.14	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu = 5$) contaminated observations when the number of variables, $p = 5$	101
Table 4.15	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu = 5$) contaminated observations when the number of variables, $p = 10$	104
Table 4.16	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu = 5$) contaminated observations when the number of variables, $p = 20$	106
Table 4.17	Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu = 5$) contaminated observations when the number of variables, $p = 30$	109
Table 4.18	Summary of appropriate clustering algorithm under normal distribution, mild and extreme contaminated observations with 10%, 20% and 30% of outliers	112
Table 4.19	Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 3$	114
Table 4.20	Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 5$	115
Table 4.21	Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 10$	116
Table 4.22	Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 20$	117
Table 4.23	Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 30$	118
Table 4.24	Summary of appropriate clustering algorithm under non-normal Gamma distribution	119

Table 4.25	Indicators for Table 4.26 until Table 4.30	120
Table 4.26	Summary of simulated results when number of variables, $p = 3$	121
Table 4.27	Summary of simulated results when number of variables, $p = 5$	122
Table 4.28	Summary of simulated results when number of variables, $p = 10$	123
Table 4.29	Summary of simulated results when number of variables, $p = 20$	124
Table 4.30	Summary of simulated results when number of variables, $p = 30$	125
Table 4.31	Computation results of multivariate normality test	127
Table 4.32	WQI classes and contamination statuses	130
Table 4.33	Correlation matrix for five water quality parameters	132
Table 4.34	Evolution indices	134
Table 4.35	Results by clusters	136
Table 4.36	List of mean centroids for each cluster from heat map analysis	139



List of Figures

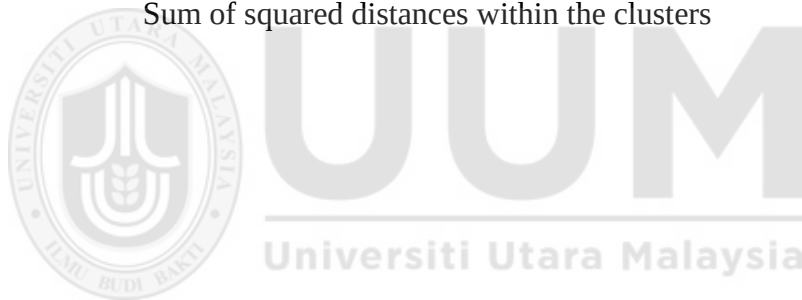
Figure 1.1	A taxonomy of clustering techniques	3
Figure 3.1	Dendrogram	54
Figure 3.2	Flowchart of the full research procedures	56
Figure 3.3	Steps of application on real data	72
Figure 4.1	Multivariate normality scatter plot	128
Figure 4.2	Spider web overview of water quality parameters	131
Figure 4.3	Agglomerative hierarchical clustering dendrogram for number of clusters is 5	135
Figure 4.4	Cluster profile plot	140



List of Abbreviations

AHC	Agglomerative Hierarchical Clustering
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
BOD	Biochemical Oxygen Demand
BSS	Average between-clusters sum of squared distances
CCC	Cophenetic Correlation Coefficient/Pekali Korelasi Kofenetik
CH index	Calinski & Harabasz index
CHAMELEON	Clustering using Dynamic Modelling
CLARANS	Clustering Large Applications based on Randomized Search
COF	Connectivity-based Outlier Factor
COD	Chemical Oxygen Demand
COR	Clustering with Outlier Removal
CURE	Clustering Using Representatives
C1	Cluster 1
C2	Cluster 2
C3	Cluster 3
C4	Cluster 4
C5	Cluster 5
DBC	Density-based Clustering
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DO	Dissolved Oxygen
DOE	Department of Environment Malaysia
ESS	Error Sum of Squares
FACADE	Fast and Automatic Clustering Approach to Data Engineering
GMM	Gaussian Mixture Model
H	Hartigan index
Hc(K)	Correction to Hartigan index
HCA	Hierarchical Clustering Analysis
LOF	Local Outlier Factor
MCD	Minimum Covariance Determinant
MVE	Minimum Volume Ellipsoid
NH ₃ -NL	Ammoniacal Nitrogen
NWQS	National Water Quality Standards for Malaysia

PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient / Pekali Korelasi Pearson
PMCC	Product Moment Correlation Coefficient
R_1 –PCA	Rotational invariant L_1 - norm PCA
ROCK	Robust Clustering using Links
RSS	Really Simple Syndication
S	Silhouette index
SOM	Self Organizing Map
SVMs	Support Vector Machines
TSS	Total Suspended Solids
UPGMA	Unweighted Pair Group Method with Arithmetic Mean / Group Average
WMCC	Weighted moment correlation coefficient
WQI	Water Quality Index
WSS	Sum of squared distances within the clusters



CHAPTER ONE

INTRODUCTION

1.1 Introduction

Clustering technique involves partitioning data with unknown groups based on their similarity. Its goals are to identify natural groupings, meaningful structures and patterns between the data points (Saxena et al., 2017; Xu & Tian, 2015). Generally, the algorithm does not know in advance either the number of clusters there should be or what the cluster should represent (Rai & Singh, 2010). This technique deals with large groups of data that have no predetermined classification from the beginning, and the observations are clustered based on their characteristics and traits under study (Sun, Chen, Xiong & Guo, 2017). Hence, clustering technique is useful in many areas such as exploratory data analysis, image and object segmentation, customer segmentation, anomaly detection, and bioinformatics (Kazi & Kurian, 2014).

Additionally, clustering technique is one of the unsupervised learning methods in machine learning where the data is not labelled, and the algorithm operates with finding patterns or structure in the data without prior knowledge of the output (Alam, Muqem & Ahmad, 2021). It means that there is no predetermined grouping that the clustering seeks to produce (Downs & Barnard, 2002). In contrast, in supervised learning, data has been labelled with dependent and independent variable(s), then the algorithm learns to predict the output based on the input variables as there is predetermined grouping of the data (Alloghani, Al-Jumeily, Mustafina, Hussain & Aljaaf, 2020; Mahesh, 2020).

Examples of methods in unsupervised learning are cluster analysis, principal component analysis (PCA), and Gaussian Mixture Model (GMM). On the other hand, examples of methods in supervised learning comprises neural networks, discriminant analysis, logistic regression, support vector machines (SVMs), Naive Bayes, and decision trees (Chauhan, Rawat, Malik & Singh, 2021; Meena & Suriya, 2020).

From time to time, clustering studies have garnered notable interest from scholars and researchers, especially for the high-dimension and multivariate data as majority of the gathered data are unlabelled or unknown in terms of their dependent and independent variables (Fahad et al., 2014; Sun et al., 2017). This technique has been massively applied in many fields of study such as social networks and information technology (Ebrahimi et al., 2022), engineering, business and marketing (Jarman, 2020), computational biology, microarray data analysis, bioinformatics, and medical imaging (Al-Jawher & Shaban, 2022; Manoharan, 2020), tourism (Jardim & Mora, 2022) and agriculture (Dange et al., 2023; Vukotic & Mircetic, 2020).

Next, there are two common approaches in clustering techniques which are hierarchical clustering and partitioning clustering. Hierarchical and partitioning approaches differ in that hierarchical methods produce a nested series of partitions, whereas partitioning methods produce only a single partition (Gulagiz & Suhap, 2017). Furthermore, a major difference between these two approaches is that partitioning methods require the number of clusters to be specified in advance, while hierarchical methods do not. As a result, hierarchical methods are often preferred over partitioning methods when no prior knowledge of the optimal number of clusters is available (Tilevik, 2017). In addition, hierarchical clustering can

accommodate arbitrary or irregular shape clusters and is suitable for high-dimensional data while the performance of partitioning clustering degrades as the data dimensionality increases and is limited to handle convex and spherical-shape clusters (Kangane, Joshi, Kacker & Jain, 2021; Nagi, Bhattacharyya & Kalita, 2011).

Based on Figure 1.1, there are two methods in hierarchical clustering which are agglomerative hierarchical clustering, also known as the bottom-up method, and divisive hierarchical clustering, also known as the top-down method.

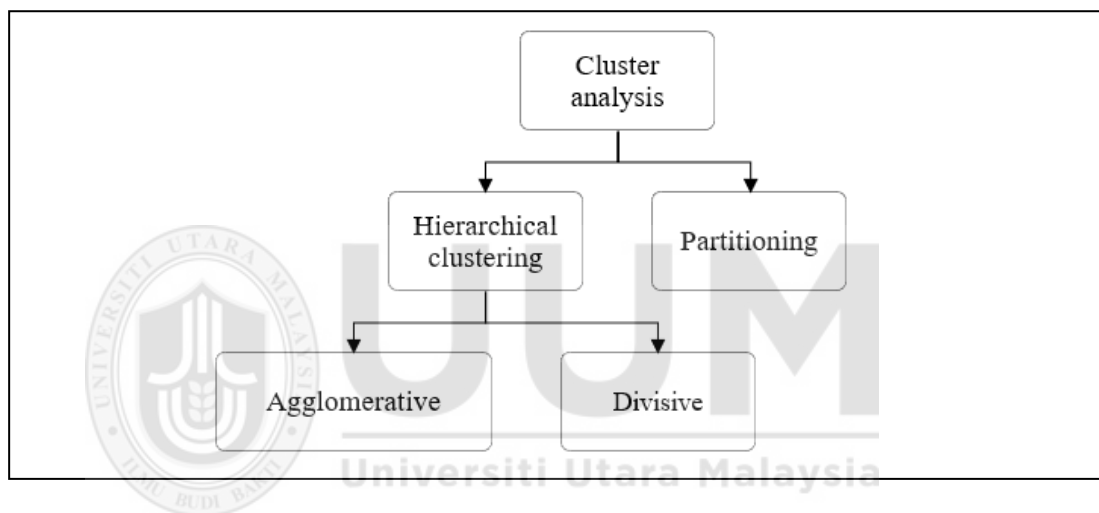


Figure 1.1 A taxonomy of clustering techniques

In agglomerative hierarchical clustering, each observation is started within its individual cluster, and couples of groups are combined as one goes up, building a hierarchy. Meanwhile in a divisive approach, initially, all data objects start in one large group and then divisions of clusters are carried out iteratively as one descends the hierarchy (Greenlaw & Kantabutra, 2014; Mohbey & Thakur, 2013). Among these two methods, agglomerative is more widely utilized rather than divisive as the divisive approach is more computational time consuming (Yim & Ramdeen, 2015). Besides being computationally costly, divisive approach is less employed owing to

the complex problem or difficulties in selecting cluster split and determining the best way to subdivide clusters (Lalitha, 2014; Shetty & Singh, 2021).

Generally, the concept of agglomerative hierarchical clustering starts with individual objects and then merges two objects or clusters into one cluster during each step resulting all objects are in one big cluster. Given n is the data points or number of observations to be grouped, first, each observation is considered as a sole independent object or cluster. The next step is to iteratively merge the most similar object or cluster creating another pair of clusters by using the similarity measure of $n \times n$ matrix between pairs of data points or objects. The similarity measure in here quantify the distance between two data points or objects which is essential for the clustering algorithm to determine which and how the object should be grouped together based on their similarity.

After that, the distance between the pairs of clusters is computed and merged by using the linkage methods reducing the number of clusters by one. Linkage method in here measured the similarity between clusters and helps to decide which clusters should be merged to create another larger cluster. All of these steps are repeated until all the observations or clusters are formed into one big cluster. Lastly, the results of the clustering technique are graphically presented in the form of a dendrogram. However, there are various types of similarity measure to represent the distance, and several choices of linkage methods to link the closest pair of clusters that can be used. Both are two crucial steps that need to be carefully considered and effectively decided when constructing a good-quality dendrogram.

There are two common approaches for determining similarity which are correlation-based distances and non-correlation-based distances. Correlation-based distances

such as Pearson Correlation Coefficient (PCC) argues that two data points with a strong correlation should have a high similarity, whereas non-correlation based distances such as Euclidean distance states that two data points with a small distance ought to have a significant similarity (Xu, Finkelman, Dai, Xu & Peng, 2021). Correlation-based distances is suitable for capturing linear and non-linear relationships among observations (Ugwu, Miasnikof & Lawryshyn, 2023). The advantage of correlation-based distances is that they are applicable to any number of random variables, rather than being limited to only two variables. It is commonly used in gene co-expression study (Chen, Ng, Lin, Zhang & Li, 2023; Flores, Inza, Larranage & Calvo, 2013), foreign exchange rate (Sharif, Yusoff & Djauhari, 2012), stocks market (Davis, Matsui, Mikosch & Wan, 2016), social sciences (Shafidan & Sharif, 2023) where relationships between variables are often of interest. It measures the closeness between two observations (Kurtzman, 2008) and also performs well for high-dimensional datasets (Shirkhorshidi, Aghabozorgi & Wah, 2015).

In general, an agglomerative hierarchical clustering algorithm is categorized into two groups of linkage methods (Murtagh & Contreras, 2012). The first group is the linkage method which includes the single linkage, complete linkage, and average linkage methods that allow a graph representation. On the other hand, the second group which allows the specification of cluster centres includes the centroid and Ward's method. In addition, single linkage, complete linkage, average linkage and Ward's method are the often-prevalent linkage methods that commonly employed in hierarchical clustering by most researchers compared to centroid method (Roux, 2018).

1.2 Issue on Agglomerative Hierarchical Clustering

In this study, we are focusing specifically on the correlation-based distances to be used as the distance measure to construct the cluster. Commonly used correlation-based distance is the PCC which had been introduced by Pearson (1896). This coefficient is usually denoted in a sample by r , and ρ for a population. It is computed on a scale with no units and can take a value range from -1 through 0 to +1. The Pearson Correlation Coefficient is a common measure of association between two continuous variables. However, there are some issues that can arise when using the PCC which are when the linearity assumption is violated, presence of outliers, and non-normal data (Armstrong, 2019; Gogtay & Thatte, 2017; Goodwin & Leech, 2006; Kozak, Krzanowski & Tartanus, 2012; Schober & Schwarte, 2018).

First, PCC assumes that the relationship between the two variables is linear. If the relationship is non-linear, the correlation coefficient may not accurately reflect the strength and direction of the association (Armstrong, 2019; Goodwin & Leech, 2006; Kozak et al., 2012). Second, PCC assumes that the data is normally distributed (Akogul & Erisoglu, 2016). If the data is non-normal, the correlation coefficient may not accurately reflect the strength and direction of the association (Armstrong, 2019; Mukaka, 2012). Lastly, PCC is sensitive to outliers, which can lead to an overestimate or underestimate of the correlation coefficient. This method is very susceptible to outliers (Kim, Kim & Ergun, 2015; Zinzendoff-Okwonu, Asaju & Arunaye, 2020) even there is a single outlier in the data, it will have an impact on the coefficient's performance (Ahad, Zakaria, Abdullah, Syed Yahaya & Yusof, 2018; Zaini & Sharif, 2020; Zakaria, Abdullah, Ahad, Yusof & Syed Yahaya, 2016).

An outlier is a data point that diverges substantially from most of the residual data. Aggarwal (2015) highlighted that while clustering aims to identify clusters of similar

data points, outliers may affect the computation of the distance that is served to categorize the objects in the clusters in which it will lead to produced significantly skewed outputs (Lopes & Gosling, 2021). Next, some researchers also refer to outliers as noise in their study. For example, Rehm, Klawon and Kruse (2007) stated that data points are deemed as noise when their distances to all clusters exceed a certain level. Consequently, the potential outliers can worsen the clustering performance (Liu, Li, Wu & Fu, 2019).

There are two alternatives for this coefficient when there is a problem of outlier(s) in the observations which are non-parametric correlation coefficient and robust approach. Non-parametric correlation coefficients such as Spearman correlation coefficient and Kendall's tau correlation are the coefficients that are suitable to use under non-normal data (Puth, Neuhauser & Ruxton, 2015). However, the computation of those coefficients involves rank values instead of the original observations which will lead to the misinterpretation of the result (Zakaria et al., 2016). Therefore, another alternative to be considered to improvise PCC when there is outlier(s) in the observations is the robust statistics. One aim of robust statistics is to deal with the data without being influenced by the outliers (Rousseeuw & Hubert, 2011). Therefore, Winsorized correlation that was introduced by Wilcox (1993) is used as the method to address the problem of PCC. Wilcox considered this estimator as a robust analogue of PCC as Winsorization is a technique that gave less weight to values in the tails and placed greater focus on values that are closer to the centre. Furthermore, Winsorized correlation is being preferred because of its ability in preventing the influence of outliers in marginal distributions of the data (Di Cesare, Luchinat, Tenori & Saccenti, 2022; Kulesz, Tian, Juranek, Fletcher & Francis, 2015; Tuğran, Kocak, Mirtagioglu, Yigit & Mendes, 2015; Wilcox, 2004). Nevertheless,

this correlation coefficient does not consider the entire data structure. As a result, a few strategically placed points can heavily impact the coefficients, masking any associations that may exist (Wilcox, 2016).

Thus, it is concluded that the alternatives and robust analogue that are being mentioned before has their own pros and cons. Therefore, it can be said that there is some difficulty in choosing a correlation-based distance measure that can satisfy agglomerative hierarchical clustering in the presence of outliers. Despite of the issue on the correlation-based distances, the selection of linkage method is also vital in producing the cluster.

Basically, single linkage defines the distance between two clusters as the minimum distance between any of their observations, while complete linkage computes the distance between two clusters as the maximum distance between any of their observations (Johnson & Wichern, 2015). Complete linkage can help to avoid chaining effect, which occurs in single linkage as it creates long chains of clusters due to the minimum distance measure it used (Bratchell, 2003; Stooksbury & Michaels, 1991). Besides, single linkage can be sensitive to outliers, while complete linkage tends to be less sensitive to outliers (Almeida, Barbosa, Pais & Formosinho, 2007). However, the weakness of complete linkage is that it has trouble when dealing with convex-shaped clusters in data because it tends to merge clusters prematurely based on maximum distances, which may not accurately represent the true structure of the data (Rani & Rohil, 2013). In addition, the average linkage was proposed to tackle the limitation of single and complete linkages (Bu, Liu, Pan & Ling, 2020).

As for the average linkage, the distance between two clusters is computed from the average distance between object's pairs where one object of a pair belongs in each

cluster. Average linkage was believed to offer more precise computation of the distance between clusters since the methodology incorporates information about the variance of the distances (Yim & Ramdeen, 2015). Prior researches showed that average linkage tends to produce more balanced cluster sizes compared to complete linkage, which encounter difficulties handling convex-shaped clusters and can break large clusters (Steinbach, Ertöz & V.Kumar, 2004; Vijaya et al., 2017) even though, complete linkage is less sensitive to outlier compared to average linkage (Carvalho, Munita & Lapolli, 2019; D'Urso & Vitale, 2020; Dabhi & Patel, 2016). Nevertheless, average linkage method is highly influenced by the shape and size of the clusters being formed. This makes it vulnerable to failure when clusters have complex shapes that deviate from the standard hyper spherical shape (Almeida et al., 2007; Guha, Rastogi & Shim, 1998; Wu & Chow, 2003). Hence, single linkage may be preferred when dealing with elongated or non-convex clusters, while average linkage may be preferred when cluster size is balance and complete linkage is favoured when sensitivity to outliers are the important factors (Saad, 2012) .

Moreover, Ward's method (Ward, 1963) also being favoured because of the ability to produce clusters that reduce variance inside-group at every dual combination (Doudney & Rinaldi, 1984). Ward's method combined the two groups that minimize the increase in the sum of squared between them (Componation & Byrd, 2000) generating clusters which help to reduce within-group variation (Murtagh & Legendre, 2014). When equated to other linkage techniques, Ward's method provides more accuracy in terms of cluster results since it minimizes the variance between the observations (Eszergar-Kiss & Caesar, 2017). The process of this method in creating a new cluster minimizes the increase in information loss that can be measured by error sum of squares (ESS). In simple words, Ward's clustering

method targets on combining two clusters in a group that minimize variance (Strauss & Von Maltitz, 2017). However, Ward's method also shares the weakness of average linkage in that it is sensitive to the shape and size of clusters. This method tends to fail when dealing with clusters that have irregular shapes or more complicated than the hyper spherical shape in which resulting wrong clusters (Almeida et al., 2007; Downs & Barnard, 2002; Guha et al., 1998; Kothari & Pitts, 1999; Wu & Chow, 2003).

Therefore, in this study, the focus is given to single linkage, complete linkage, average linkage, and Ward's method, since each of them has its own strengths and can be effective in different situations.

1.3 Research Problem Statement

The motivation behind this research aroused based on the previous issues regarding the sensitivity of the agglomerative algorithm to the choice of the distance measure and linkage method for cluster formation. Choosing the appropriate distance measure and linkage method is crucial, as different selections can potentially yield different results, in which can affect the interpretation of the clusters. The wrong selection of the distance measure and linkage method may cause the agglomerative algorithm to suffer from the chaining problem, where two clusters are merged too early in the process, leading to the formation of long and thin clusters that may not be meaningful (Kaufman & Rousseeuw, 1990; Saracli, Dogan & Dogan, 2013). Additionally, a good combination of distance measure and appropriate linkage technique in producing the clusters will help the researcher to produce the best segmentation (Choudhury & Ashraf, 2022).

Second problem arises when there is an outlier in the data, especially when using PCC, which is susceptible to outlier(s). Empirically, outliers in the dataset can lower the correlation value and may substitute the sign from negative to positive and vice versa (Zinzenoff-Okwonu et al., 2020). The root cause of this problem is because of the usage of mean in the calculation. According to Zakaria et al. (2016), the mean has a zero-breakdown point and the influence function is limitless. Additionally, the non-parametric methods, Spearman's correlation and Kendall's Tau coefficients have the weakness of potentially leading to misinterpretation of the result, as they are using rank values instead of the original observations in their computation. Meanwhile, the Winsorized correlation, an alternative to the PCC, has the disadvantage of masking data, as it does not consider the entire data structure. Therefore, selecting a correlation-based distance measure that can effectively supports agglomerative hierarchical clustering in the presence of outliers has been very challenging. Furthermore, if the outliers have significant impact, the researcher may not be able to identify the distinction of the clusters or that their objects are internally homogenous. As the outliers are the values that deviate from the expected range of data variation, it may affect the computation of the distance that is served to categorize the objects in the clusters. Thus, it will lead to produced significantly skewed outputs (Lopes & Gosling, 2021).

Moreover, various modifications and extensions to the agglomerative algorithm have been proposed by previous studies in order to improve the quality of clustering performance, such as using approximate methods for distance calculations (Choudhury & Ashraf, 2022; Prasetyo & Purwarianti, 2015) pruning the dendrogram to remove uninformative clusters (Ge & Tibshirani, 2022) and incorporating prior knowledge to guide the clustering (Ma & Dhavala, 2018). However, there are no

proper guidelines have been provided from the previous studies to the users or researchers. Researchers face difficulties in selecting the appropriate clustering algorithm for their real data applications. Supposing a table of summary can be referred as a guideline on which distance measure and linkage method perform well at specific sample size and number of variables, this is the third problem in this study.

These three obstacles as explained above have motivated us to further explore the statistical performance on clustering algorithms to help the researcher in choosing the appropriate distance and linkage methods using table of summary. To address the problems, it is also necessary to provide reliable results on the appropriate clustering technique to be used in the presence of outlier in the data. Therefore, to solve this problem, the evaluation of the statistical performance on sixteen (16) clustering algorithms produced from four (4) correlation-based distance measures and four (4) linkage methods under different conditions of observation is required. The performance evaluation is based on the calculation of the average value of Cophenetic Correlation Coefficient (CCC). The Cophenetic Correlation Coefficient (CCC) measures how faithfully a dendrogram preserves the pairwise distances between the original, unmodeled data points (Rohlf & Fisher, 1968). From those results, a table of summary is produced. A researcher will have a recommendation on which correlation-based distance and linkage are appropriate at certain number of variables, sample sizes and different distribution. Besides, the illustration on real data of Juru river basin water quality also being performed to show the application of appropriate clustering model and its interpretation to show the advantage of the clustering technique. For this purpose, the CCC and Calinski & Harabasz index (CH) are used to evaluate the performance of the clustering model.

1.4 Research Questions

The research questions are as follows,

1. How is the performance of the 16 clustering algorithms constructed from of four (4) correlation-based distances and four (4) linkage methods?
2. Which correlation-based distance is appropriate when there is an outlier in data?
3. How can we summarize the results of statistical performance in a convenient way to the researcher?
4. How can we illustrate the application of the appropriate clustering model on real data?

1.5 Research Objectives

The main objective of this study is to measure a statistical performance of agglomerative hierarchical clustering algorithms that are formed from several correlation-based distances and several linkage methods using simulation study and real data. The sub-objectives of the study are as follows:

1. To investigate the performance of 16 clustering algorithms performance formed from four (4) different correlation-based distances and four (4) different linkages methods by calculating the average value of CCC.
2. To determine the appropriate correlation-based distances based on CCC when there are outliers in data under several simulated condition.
3. To propose a table of summary as a guideline for choosing appropriate clustering algorithm to the researcher.
4. To verify and demonstrate the appropriate clustering model for Juru River Basin water quality real dataset.

1.6 Research Significant

This study aims to provide several contributions to various parties. Firstly, it seeks to contribute to the body of knowledge, especially in the development of robustness theory within the field of clustering. The significance of this study lies in its aim to offer insights and knowledge regarding which clustering algorithm is suitable to be used based on the types of data distribution and when there is presence of outliers in the data.

Furthermore, the novelty of this research is its exclusively focus on correlation-based distance measures used to construct clustering algorithms. Thus, future researchers can refer to this study as a guideline to gain information on the appropriate correlation-based distance measures and also linkage methods for constructing clustering algorithms for various types of data distributions, especially when applying clustering techniques to real data in their respective fields of expertise.

Moreover, this study has the potential to assist stakeholders and government entities, such as the Penang State Government and the Department of Environment Malaysia, in addressing the crisis of contamination and water quality in the Juru river basin. The clustering algorithm may be beneficial and applicable for assessing and analysing water quality parameter based on the Water Quality Index (WQI), enabling them to identify the level of river contamination and take necessary measures to address the pollution issue. The results can also aid the Malaysian government in formulating better policies and in educating the community about river contamination.

In addition, this study also valuable to the community, including industry players, workers, and residents in the Juru area. It can provide information on the level of

river contamination and, to some extent, the substances responsible for river pollution. This information can raise awareness within the community and encourage actions to protect the river, such as controlling the discharge of domestic sewage into the river.

1.7 Research Scope

This study has several research scopes. The central focus of this study is on agglomerative hierarchical clustering algorithms, with a specific emphasis on the utilization of the correlation-based distance measures. The simulation study conducted to evaluate the statistical performance of 16 clustering algorithms is based on the CCC under various simulated conditions. The 16 clustering algorithms that are produced is from the pairing of only four types of correlation-based distances i.e., classical PCC, Spearman correlation, Kendall's Tau correlation and Winsorized correlation coefficients with four different linkage methods which are average linkage, Ward's method, complete linkage and single linkage. Additionally, the presence of outliers in the data is also addressed in the simulation study. Lastly, the clustering model is applied only on real Juru river water quality parameter dataset taken from Department of Environment Malaysia (DOE), which encompasses data collected over the past 10 years, from 2008 until 2017. The CCC and CH indices are used to evaluate and validate the performance of the clustering model after its application to this real dataset.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This study focuses on hierarchical clustering method because it provides detailed information and comprehensive view of the data structure via the dendrogram. The agglomerative hierarchical clustering is used as it is more computationally efficient and has a straightforward process that is easier to code and understand which begin with each data point as a separate cluster and iteratively merge the closest pairs of clusters. In addition, the primary point of interest in this study is the correlation-based distances which can capture the behaviour of dataset and is suitable for high dimensional dataset. Therefore, in this chapter, it covered the review and discussion on past literature and research on hierarchical clustering, distance measure, outliers, non-parametric methods, robust estimator, linkage method and clustering validation measures.

2.2 Hierarchical clustering

Clustering or also known as cluster analysis involves utilizing statistical techniques to classify objects or data points into distinct clusters or segments based on their shared attributes or characteristics. This method is widely applied in statistical analysis, data mining, and machine learning. The ultimate objective of cluster analysis is to create meaningful groups that can be further explained and analysed (Rodriguez et al., 2019) or used in decision-making processes. The process of cluster analysis entails identifying similarities and differences among data points and grouping them accordingly, based on their degree of similarity.

There are two main advantageous of using the clustering techniques that had been highlighted in many studies. First, clustering helps in exploring and understanding

the underlying patterns and structures within the data without the need for pre-existing knowledge about the classes or categories in the data. It can reveal insights about relationships, similarities, and differences among data points or observations that might not be apparent initially (Arora & Chana, 2014; Kazi & Kurian, 2014; Pei, Zhang, Cho, Wang & Yu, 2003; Sun et al., 2017; Wang, Wang, Yang & Yu, 2002). Secondly, as high-dimensional data can be difficult to visualize and analysed, clustering is used to help reduce the dimensionality of the data by grouping similar features together, making it more manageable and easier to identify redundant or similar data points in a dataset. This can aid in data cleaning and preprocessing, leading to improved model performance (Alam et al., 2021; Fahad et al., 2014; Nguyen et al., 2019; Pandove, Goel & Rani, 2018; Sajana, Sheela Rani & Narayana, 2016).

Therefore, based on the aforementioned advantages, numerous clustering algorithms such as partitional, hierarchical, density-based clustering (DBC) methods and others have been extensively developed, studied and applied by researcher in various fields for over the last few decades. One of the old studies on cluster analysis is by Scoltock (1982) where he proposed that cluster analysis have been implemented on data presented in five major groups: biology and zoology, medicine and psychiatry, sociology, criminology, anthropology and archaeology, geology, geography and remote sensing, and the fifth group is on information retrieval, pattern recognition, market research and economics. Badr, Zaitchik and Dezfuli (2015), Fovell (1997) and, Stooksbury and Michaels (1991) employed cluster analysis in the study of climate zones. Next, clustering also has been used for segmentation in business marketing (Arguelles, Benavides & Fernandez, 2014; Lizbetinova, Starchon, Lorincova, Weberova & Prusa, 2019) where the tools for the purposed of marketing

communication used by enterprises are being clustered. This had aid businesses in tailoring their marketing strategies to have specific segments and addressing their unique needs. Componation and Byrd (2000) has applied clustering as a feature engineering step. Instead of using the raw data, they used cluster labels as features, which capture more relevant information in order to establish a systematic approach for concurrent engineering teams, enabling projects to be completed more faster and with a reduced likelihood of project failure. Additionally, clustering has also been widely utilized in biomedical research to identify groups of genes with similar expression patterns in gene expression data (Babichev, Taif, Lytvynenko & Osypenko, 2017; Chen et al., 2023; De Souto, Costa, De Araujo, Ludermir & Schliep, 2008; Patrik, 2005; Vidman, Kallberg & Ryden, 2019; R. Xu & Wunsch, 2010) as it helps the researchers understand gene functions and their roles in biological processes.

In addition, water quality data has also been a subject of interest in numerous studies that have employed cluster analysis. For instance, Aubert et al. (2013) utilized cluster analysis to identify flood patterns in multivariate time series water quality data gathered from the Kervidy-Naizin watershed. Similarly, Lyra, Oliveira-Junior and Zeri (2014) applied cluster analysis to rainfall data with the aim of identifying rainfall patterns based on month and year. Lee et al. (2020) and Warsito, Sumiyati, Yasin and Faridah (2021) used cluster analysis to study and monitor the river water quality dataset. However, since cluster analysis categorizes objects into clusters based on their similarity or dissimilarity, the results obtained can vary considerably depending on the type of distance measure used (Lee et al., 2020). Hence, in this study, we also did utilize cluster analysis to group and analyse the water quality parameters of Juru River Basin dataset.

Next, there are two common types of clustering techniques which are hierarchical clustering and non-hierarchical clustering or partitioning technique (Faroh, Tajuddien, Narimawati, Affandi & Priadana, 2022). Hierarchical clustering technique or also known as hierarchical cluster analysis (HCA) involve generating clusters where object with similar characteristics is group together where the actual group of every object is unknown (Hamzah, Kek & Saharan, 2017). The clusters' hierarchy in HCA is demonstrated by the division and combination of clusters at separate levels, meanwhile, in partitioning technique, clusters are determined directly by detailing the number of clusters, K from the beginning. Generally, hierarchical clustering techniques is divided into two classifications which are a number of consecutive mergers or knowns as agglomerative clustering, and a series of succeeding divisions that known as divisive clustering. Both are also called as bottom-up and top-down, respectively. Meanwhile, non-hierarchical clustering techniques begin with either a preliminary partition of objects into clusters that produce the clusters' nuclei (Saxena et al., 2017).

However, one of the issues in clustering is there is no ideal algorithm that can be applied to produce the ultimate data segmentation (Jarman, 2020). This limitation is faced by researchers because there is no measurement to indicate the best clustering algorithm. Numerous investigations and mathematical aspects are required before choosing an effective clustering algorithm either hierarchical or non-hierarchical clustering (Jarman, 2020). According to Gulagiz and Suhap (2017) when it is compared with hierarchical clustering technique, non-hierarchical techniques is low cost in terms of computation time. Furthermore, Oyelade et al. (2019) also specified that partitional clustering technique stays as one of the most popular and applied techniques due to its simplicity, efficacy and easy to employ. However, in terms of

information and description of the clustering structure, hierarchical techniques were said to provide richer explanation and picture of the cluster structure than the non-hierarchical methods (Campello, Moulavi, Zimek & Sander, 2015; Chaakraborty, Sen & Maity, 2020) as it describes data from numerous levels of pairing of clusters. Additionally, Kimes, Liu, Neil Hayes and Marron (2017) has highlighted that hierarchical clustering delivers a spontaneous way to investigate the associations between clusters as it went along the stages of producing a single nested hierarchy of clusters. Hierarchical clustering also capable of handling clusters with arbitrary or irregular shapes and is well-suited for high-dimensional data, whereas partitioning clustering experiences trouble in clustering data when the dimensionality of the increases and have various size and density (Bindra, Mishra & Suryakant, 2019; Nies et al., 2019). Moreover, dendrograms, an exceedingly informative depiction of hierarchical clustering as a binary tree may also have a role in making hierarchical clustering technique popular in practice. Dendrogram is a graph that visualize the topological of dataset (Zulkepli, Md Noorani, Razak, Ismail & Alias, 2020). As a conclusion, in this study, we will further investigate about the hierarchical clustering technique.

Moreover, HCA technique is divided into two types of approaches which are divisive and agglomerative approaches. The techniques that begin with a single big cluster or group, and divide it into several clusters are referred as divisive method, whereas procedures that begin with clusters containing a single point and then merge them as one cluster are referred as agglomerative method (Yusniyanti, Virgantari & Faridhan, 2021; Zhou, Xu & Liu, 2017). Another definition is also given by Wijuniamurti, Nugroho and Rachmawati (2022), the procedure of agglomerative approach begins with each object in a separate group and then consecutively merges the groups,

decreasing the number of groups at each level until all objects belong to only one group, whereas the divisive strategy begins with all of the objects in one group and then divides them into smaller groups until there is only one object per group.

Rajalingam and Ranjini (2011) in their work had compared the performance of agglomerative and divisive method in terms of running time computation of the algorithm. They had found that the computational of divisive method is twice faster than agglomerative method. On the other hand, Roux (2018) concluded that divisive hierarchical algorithm is time consuming because it involved dividing the cluster into two sub-clusters, and again divided into sub-clusters, and so on which produce $2^{n-1} - 1$ ways of dividing of a set of n objects. The larger the n , the number of ways will be tedious. Yet, Blashfield and Aldenderfer (1988) had stated that agglomerative method is the most commonly applied clustering method because of the computational simplicity and produce non-overlapping clusters. Thus, this method was said to yield better quality of clusters and it can be further used to generate taxonomy (Vijaya et al., 2017). Apart from that, studies from Nisha and Kaur (2016), Mercioni and Holban (2018) give us the motivation to bring the focus into agglomerative algorithm. Nisha and Kaur (2016) summarized that agglomerative algorithm is superior than divisive algorithm in real practice. Apart from that, a major downside of divisive algorithm is the difficulty in deciding the next cluster to divide, where usually it will be based on the size of the cluster. The algorithm may only choose to divide the largest cluster instead of the smaller one (Mercioni & Holban, 2018).

To conclude, clustering is a technique used to classify objects or data points into different groups based on their similarity features, and it has been a central theme for this study. This technique had been widely applied across various field due to its advantages on helping to discover underlying patterns and structures in data without the need for pre-existing knowledge of categories, aiding in managing high-dimensional data and simplifying data analysis. In this study, hierarchical clustering was preferred as it offers richer understanding of cluster structures. The agglomerative method which is noted for its computational simplicity and better cluster quality also had been used. In the agglomerative clustering algorithm, the distance measure plays a crucial role. The next section will review the distance measure and its various types.

2.3 Distance measure

As stated, before in Chapter 1, agglomerative clustering needs similarity measures to build hierarchy of clusters whereby similar nature of observations or clusters are merged into same groups forming few numbers of clusters and the process continues until all clusters belong to only last one cluster. Commonly, similarity is described in terms of distance measures between the observations (Tilevik, 2017). Distance measures is an important tool in clustering because the algorithm consists of distance between two objects, and also distance between two clusters in the dataset (Hadi, 2022). In this section, we will briefly elaborate the similarity or distance measures specifically on correlation-based distance.

Similarity by a simple meaning is the state of being almost the same. In statistic, similarity measure provides the concept of proximity or closeness in which it presents ‘how close’ or ‘how far’ the observations are to each other. In clustering,

similarity measure is a fundamental aspect because observations are clustered based on their similarity in which similar observations are grouped into the same cluster while dissimilar observations are assorted into different cluster (Torres, Basnet, Sung, Mukkamala & Ribeiro, 2009).

Many clustering algorithm approaches apply distance measures to determine the similarity or dissimilarity of any pair of objects. The improper usage of distance and dissimilarity measures on a clustering method might result in poor clustering results (Prasetyo & Purwarianti, 2015). Rajalingam and Ranjini (2011) also highlighted that one of the significant algorithms in clustering is the identification of similarity or distance measure which will discover how the proximity of two objects or items is computed. This will affect the formation of the groups as some objects might be near with each other corresponding to one distance measure and some may be far apart from each other, depending on the usage or type of distance measure.

In addition, various studies had been done in comparing and emphasizing the usage of distance or similarity measure in building clusters. Kumar, Chhabra and Kumar (2014) examined how the performance of clustering algorithms is influenced by different distance measures. They evaluated the performance of six distance measures which are Euclidean, Euclidean Squared, Manhattan, Mahalanobis, Cosine similarity, and Pearson correlation using three clustering algorithms, K-means, single linkage and average linkage on artificial and real datasets. The authors highlighted that the choice of distance measure can have a significant impact on the clustering results and that the effectiveness of each measure depends on the data type. Therefore, the authors recommended that researchers should take into account the characteristics of their data and the objectives of their analysis when selecting a distance measure and a clustering algorithm.

Furthermore, Shirkhorshidi et al. (2015) in their article had discussed a comparison study on different similarity and dissimilarity measures used in clustering continuous data. The authors examined twelve (12) distance measures, including Euclidean distance, Pearson correlation, cosine similarity, and evaluate their performance in terms of accuracy and efficiency. The study asserted that the choice of similarity measure can significantly affect the clustering results, and suggests that researchers should carefully consider the characteristics of their data and the goals of their analysis when selecting a measure.

Previously mentioned, the similarity in clustering is commonly calculated by using distance measure. There are two types of distance measures and one of them is distance-based measure such as Euclidean distance, Maximum, Manhattan, Minkowski and Chebyshev distance wherein large distance depicted low similarity and vice versa. This distance metrics is usually used in measuring the similarity or dissimilarity of continuous variables (Dahal, 2015). However, the limitation of this type of distance measure is that it computes only the physical distance between two data points and thus, are insufficient when it pertains in describing the behaviour of the dataset (Irani, Pise & Phatak, 2016). In addition, this type of distance metric such as Euclidean distance does not take into consideration the correlation between variables and may not totally suitable for high dimensional datasets (Ertoz, Steinbach & Kumar, 2003; Mimmack, Mason & Galpin, 2001).

Therefore, another frequent technique for identifying similarities between items is to use distances based on correlation such as Pearson Correlation Coefficient (PCC) in which we were focusing in this study. The correlation-based distance measures consider two objects as similar if their attributes show substantial correlation, even if their actual values might be distant when measured using Euclidean distance

(Kassambara, 2015). As for this form of distance measure, it uses correlation coefficient to determine the correlation between the items where two items with a strong positive correlation coefficient will have a distance close to one which indicates a high similarity between the items (Gibbons & Roth, 2002; Xu, Peng, Li & Xu, 2020). Hence, by using this type of distance measure, the behaviour of the dataset can be captured and the probability of the existence of similar data patterns among a set of data points can be detected even if they are far away from each other (Kulkarni, Tokekar & Kulkarni, 2015).

In this study the correlation-based distances that are being studied are PCC, Spearman correlation, Kendall's Tau correlation and Winsorized correlation upon which it'll be described further in the next sections respectively.

2.4 Parametric Method: Pearson Correlation Coefficient (PCC)

Pearson Correlation Coefficient (PCC) or also expressed as Product Moment Correlation Coefficient (PMCC) has been introduced and named after Pearson (1896). This coefficient aimed to calculate the strength of a linear relationship between two variables. Its value lies from -1 through 0 to +1. +1 indicates that there is a perfect positive relationship or strong correlation between the variables and vice versa. In addition, an increase in the first variable is associated with a corresponding increase in the second variable, demonstrating a positive correlation coefficient that signified a direct relationship between the variables. Conversely, a negative correlation suggests an inverse relationship where an increase in one variable results in a decrease in the second variable. In contrast, when the correlation coefficient is 0, there is no discernible relationship between the two variables (Taylor, 1990).

Next, PCC was categorized as parametric method that refers to a type of statistical technique that assumes the data has a normal distribution (Kassambara, 2015). PCC relies on several assumptions regarding the data, including a linear relationship between the two variables, normal distribution, and no substantial outliers (Armstrong, 2019; Gogtay & Thatte, 2017; Goodwin & Leech, 2006; Kozak et al., 2012; Li & Kim, 2003; Schober & Schwarte, 2018). Goodwin and Leech (2006), Kozak et al. (2012) and Armstrong (2019) asserted that if the relationship between the variables is non-linear, the strength and direction of the association may not be reflected precisely by the correlation coefficient. Additionally, if the data is not normally distributed, the correlation coefficients should be calculated from the ranks of the data rather than their actual values wherein non-parametric methods such as Spearman correlation should be used (Akoglu, 2018; Alsaqr, 2021; Armstrong, 2019). Finally, the presence of outliers can affect PCC's performance and lead to misinterpretation of the coefficient (Gogtay & Thatte, 2017; Y. Kim et al., 2015; Kozak et al., 2012; Mukaka, 2012; Schober & Schwarte, 2018).

Nevertheless, the use of PCC in clustering has also been extensive in various fields. For instance, Hu and Pan (2015) employed PCC as the distance measure in hierarchical clustering algorithm to cluster the words and contents of *Really Simple Syndication* (RSS) datasets. Xu et al. (2021) in their study used the average linkage hierarchical clustering algorithm with PCC as the similarity measure for the analysis of the geological processes while Strickert, Sreenivasulu, Villmann and Hammer (2008) utilized PCC with centroid-based clustering to cluster gene expression data and mass spectra. Blenkiron et al. (2007) make use of PCC with average linkage method to clusters the data on human breast tumours. Despite being utilized in various fields, PCC has also faced criticisms whenever the assumptions are violated.

For examples, Kurtzman (2008) highlighted that the use of PCC as a distance metric is not appropriate for clustering microarray data due to its susceptibility to outliers. He then suggested Tukey's biweight metric as the metric for clustering microarray data in his study. Kassambara (2015) advised to use Spearman correlation instead of PCC when clustering genes as PCC is vulnerable to outliers.

2.5 Outliers

Many researchers defined outliers in numerous ways depending on the area of application. In the literature of data mining and statistics, outliers are also referred as abnormalities, discordant, noise or anomalies (Ankur & Singh, 2014; Panjei, Gruenwald, Leal, Nguyen & Silvia, 2022). Hawkins (1980) referred outlier as an observation that differs so significantly from the other observations that it raises the doubt that it was produced by another mechanism. In simple words, data point that deviates significantly from most of the remaining data is denoted as an outlier. Nonetheless, an outlier in a dataset can also sometimes deliver a vital message regarding uncommon but important circumstances that are not present in majority of the data. In such cases, the outlier serves as a signal to further investigate why it differs from the remainder of the data. However, in another circumstances, the outlier may exist as a result of data entry errors (Kwak & Kim, 2017). If the purpose of the analysis is to identify the central patterns as reflected by the bulk of the data, then an outlier might have more significant impact on the results obtained rather than the effect of any other single data point (Alameddine, Kenney, Gosnell & Reckhow, 2010).

According to Aggarwal (2015), while clustering aimed to identify groups of similar data points, outliers were unique data points that vary from majority of the data.

Rehm et al. (2007) stated that data points whose distances to all clusters exceeded a certain threshold were considered as outliers. There are numerous studies had been conducted over the past years regarding the issue on outliers in clustering specifically on how to handle outliers or noisy data. Mangiameli, Chen and West (1996) highlighted that even though hierarchical clustering method perform well for compact datasets, it failed to work well when there was presence of outliers. Thus, in their study, they proposed Self Organizing Map (SOM) network as the alternative and had compared the performance of the SOM network to single, complete, average linkage, centroid method, Ward's method, two stage density and *Kth* nearest neighbour clustering methods. From their study, it is founded that SOM network was more robust and superior than the other seven clustering methods in determining cluster membership even when outliers are present in the data.

In addition, several robust clustering techniques have also been introduced and suggested to handle the outlier's issue over the past decades. For example, Zhang, Ramakrishnan and Livny (1996) presented Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), that fitted for very large databases and also the first clustering algorithm that was being introduced in the area of handling database that contain outliers. Guha et al. (1998) had introduced Clustering Using Representatives (CURE) which employed hierarchical clustering algorithm that adopted a compromise position between the centroid-based and all-point extremes. Qian, Zhang and Zhang (2004) proposed Fast and Automatic Clustering Approach to Data Engineering (FACADE), a novel hierarchical clustering technique that was capable of effectively processing large volumes of spatial data containing significant noise, thereby accurately identifying natural clusters. Ding, Zhou, He and Zha (2006) proposed robust Principal component analysis (PCA) clustering known as rotational

invariant L_1 - norm PCA ($R_1 - PCA$) to reduce the undesirable effect of outliers in the PCA cluster structure. Next, Almeida et al. (2007) introduced an improved single linkage hierarchical clustering that is based on self-consistent outlier reduction approach and followed by the construction of a descriptive function. Liu et al. (2019) proposed Clustering with Outlier Removal (COR) algorithm that considered the joint clustering and outlier detection problem. In addition, instead of these robust clustering techniques being proposed by those researchers, they were also being evaluated and compared in terms of their performance with other clustering methods in the studies, respectively. Summary of these past studies is expressed in the Table 2.1.

Table 2.1

Past studies on the issue on outliers in clustering

Authors and Years	Issues	Methods involved and being compared	Findings
Mangiameli et al. (1996)	Conventional hierarchical clustering is not robust to outliers	SOM network, single linkage, complete linkage, average linkage, centroid method, Ward's method, two stage density, and Kth nearest neighbor	SOM network is superior than others
Zhang et al. (1996)	Prior clustering method is not effectively addressed the problem of large datasets and outliers	BIRCH, Clustering Large Applications based on Randomized Search (CLARANS)	BIRCH outperformed CLARANS
Guha et al. (1998)	Existing clustering techniques are fragile to outliers	CURE, BIRCH	CURE outperformed existing clustering algorithms and can perform well for large databases

Qian et al. (2004)	Hierarchical clustering analysis (HCA) is not robust to outliers	FACADE, Clustering using Dynamic Modelling (CHAMELEON)	FACADE perform better
Almeida et al. (2007)	HCA sensitive to outliers	Improved agglomerative HCA based on single linkage, Robust Clustering using Links (ROCK), CURE, Density Based Spatial Clustering of Applications with Noise (DBSCAN), CHAMELEON and FACADE	The improved version of the single linkage is the most efficient method
Ding et al. (2006)	PCA is sensitive to outliers	R_1 –PCA, PCA	R_1 – PCA is effective in handling outliers
Liu et al. (2019)	K-means cluster structure is vulnerable to outliers	K-means, K-means--, COR	COR is efficient
Nowak-Brzezinska and Gaibei (2022)	Data containing outliers can hinder in producing good and well separated clusters	Agglomerative Hierarchical Clustering (AHC), K-means, Local Outlier Factor (LOF), Connectivity-based Outlier Factor (COF)	COF better than LOF in removing outliers. AHC is more resistant to outliers than K-means

Similarly, in the context of PCC, ample studies have been done regarding the outlier problems in previous years. This study also had highlighted this issue in order to determine which correlation-based distances is appropriate to be used when there are outliers in the data. In the realm of focus, another approaches that has been discovered and used by researchers to overcome the issue of outliers is non-

parametric method and robust approach that will be describe further in the next sections of this chapter.

2.6 Non-parametric methods

Non-parametric methods are statistical methods that do not relies on assumptions wherein the data used can be acquired from a sample that does not fit into any particular distribution such as normal distribution. Unlike parametric methods such as PCC, which often assume that the data follows a particular distribution (such as normal distribution), non-parametric methods do not rely on these assumptions and are more flexible in their application (Colwell & Gillet, 1982). They are two common correlation coefficients under non-parametric methods which are Spearman correlation coefficient and Kendall's Tau correlation coefficient (Puth et al., 2015). According to Croux and Dehon (2010) in their study, they asserted that the influence function of these non-parametric correlation measures is bounded, confirming their robustness to outliers. Despite their robustness, they maintain a desirable level of statistical efficiency. While the Pearson correlation is the most efficient under normal distribution, the Spearman and Kendall correlation coefficients maintain over 70% statistical efficiency across all population correlation values, striking a balance between robustness and efficiency.

2.6.1 Spearman Correlation Coefficient

Spearman correlation coefficient which has been introduced by Spearman (1904), is identical to PCC, but it is calculated using the ranks of the data instead of the original raw data. It measures how well the relationship between two variables can be defined via a monotonic function. A monotonic function refers to the type of relationship between two variables where the relative order of their values remains consistent, but

not necessarily linear (Raveh, 1986). When a Spearman correlation is positive, it indicates that higher values of one variable tend to correspond to higher values of the other variable in a consistent manner, even if the relationship isn't strictly linear. Similarly, a negative Spearman correlation suggests that higher values of one variable tend to correspond to lower values of the other variable in a consistent way (Sedgwick, 2014). In contrast with monotonic function, a linear function which is used in the context of PCC, is a specific type of relationship where the change in one variable is directly proportional to the change in the other variable (Heuvel & Zhan, 2022). In other words, the relationship follows a straight line on a scatter plot.

Spearman correlation coefficient ranges from -1 to +1, where the absolute value of the coefficient indicates the strength of the monotonic relationship. A value of -1 indicates a perfect negative monotonic relationship and +1 indicates a perfect positive monotonic relationship. A value closer to 0 signifies a weaker monotonic relationship between the variables. Nevertheless, just like PCC, Spearman's coefficient can be 0 for non-monotonic relationships. Unlike PCC, Spearman's coefficient can be 1 for both linearly related variables and variables related by some non-linear but monotonic manner (Chok, 2008).

Moreover, the advantage of Spearman over PCC is that it is robust to outliers. As the actual values of the sequences are substituted by their ranks, Spearman tends to be less affected by outliers compared to PCC (Janse et al., 2021; Jaskowiak, Campello & Costa, 2014). Additionally, Spearman calculates the correlation coefficient by comparing the rank orders of the two variables, making it suitable for cases where the relationship may not be linear or where the data doesn't meet the assumptions of normality (Alsaqr, 2021; Zar, 2005).

2.6.2 Kendall's Tau correlation

Kendall's tau correlation is another non-parametric method. It is a rank-based correlation coefficient that was introduced by Kendall (1938). Similar to Spearman's correlation, Kendall's tau assesses the strength and direction of a monotonic relationship between two variables, but it focuses on the number of concordant and discordant pairs of data points (Croux & Dehon, 2010). For each pair of data points, the concordance or discordance of their ranks is determined. Two pairs of data points are concordant if their ranks have the same order (either both increasing or both decreasing), and they are discordant if their ranks have opposite orders (Newson, 2002).

Just like the two aforementioned correlations, Kendall's Tau also lie on a scale from -1 to +1, where the values near -1 signify negative relationship, near to +1 signify positive relationship and 0 indicates no ordinal association at all (Chok, 2008).

Furthermore, similar to Spearman correlation, Kendall's Tau is particularly useful when capturing non-linear monotonic relationships between variables and dealing with ordinal or ranked data (Heuvel & Zhan, 2022; Puth et al., 2015). Additionally, Kendall's Tau is also robust against outliers, meaning that extreme values in the data are less likely to significantly affect the correlation value compared to PCC (Jaskowiak et al., 2014; Rajesh & Chaturvedi, 2019).

2.7 Winsorized Correlation Robust Estimator

Robust estimators are estimators derived for a given model while taking into account minor deviations from this model. One of the fundamental principles of robustness is that a robust estimator should be unaffected by a small amount of contamination regardless of its location (Filzmoser, Serneels, Maronna & Van Espen, 2009). There

are two different approaches that can be applied for robust estimation procedures. First, the detection and elimination of outliers will be made before applying traditional estimators to the residual good data points. Second, the classical estimators are being substitute with robust estimators (Beckman & Cook, 1983). For instance, trimmed mean and Winsorized variances were applied as the central tendency measure and scale estimator, respectively in the Welch test, as described in the study conducted by Kowalchuk, Keselman, Wilcox and Algina (2006).

There are many types of robust estimators that have been introduced and debated in the past researches such as Minimum Covariance Determinant (MCD), Minimum Volume Ellipsoid (MVE), trimmed mean and many more. As for the robust estimators that is associated with PCC, one of them is Winsorized correlation coefficient, in which is being used in this study.

As the performance of Pearson's correlation is affected by outliers, few alternatives of PCC that have been introduced by several researchers are Spearman correlation (Spearman, 1904) and Kendall's Tau correlation coefficient (Kendall, 1938). However, the disadvantage of these coefficients was the usage of rank values instead of original observations which may lead to misinterpretation of the result (Zakaria et al., 2016). Thus, another alternative of measures of PCC is Winsorized correlation.

Another robust analogue of PCC for the issue of the presence of outliers and non-normal data is Winsorized correlation which had been introduced by Wilcox (1993). Hastings, Mosteller, Tukey and Winsor (1947) discovered winsorization as a transformation technique in data distribution by preventing the extreme values for the purpose of minimizing the effect of outliers in the data distribution. In addition, this technique focuses around the data centre rather than both sides' tails in order to

prevent bias. Fundamentally, the procedure of winsorizing is quite similar to trimming procedure, but the data that are aimed to be discarded will be swapped with the highest and lowest end of the remaining data, respectively. As result, the size of sample data is remaining the same and being preserved which prevents the loss of information. Moreover, this approach was also found to control the rate of Type I error ~~rate~~ even the data having skewed distribution (Ochuko, Abdullah, Zain & Syed Yahaya, 2016; Xao, Syed Yahaya, Abdullah & Md Yusof, 2014).

Furthermore, Kulesz et al. (2015) highlighted that the usage of Winsorized means and standard deviations in Winsorized correlation are more resilient against anomalies than sample mean and standard deviation. This is because it trims a specific proportion of the data from the tails and replace those values with less extreme values. This results in less weight is being assigned to values close to the distribution's tails and more weight is assigned to values close to the distribution's centre. Besides, it's computational is also simple and easier to be used (Mair & Wilcox, 2019).

2.8 Previous studies on the comparison of correlation coefficients

Additionally, many past researches have been conducted to compare all of these methods to see which of the correlation coefficients under each method perform better, specifically when there is a presence of outliers. The summary on the comparison of the correlation coefficients is depicted in Table 2.2 as follow.

Table 2.2

Summary on the comparison of correlation coefficients

Authors and Years	Issues	Methods involved	Findings
Wilcox (1993)	PCC susceptible to outliers	PCC, Kendall's Tau, Spearman, Winsorized correlation	Winsorized performed better

Croux and Dehon (2010)	PCC affected by outlier	Spearman, Kendall's Tau, Quadrant and Pearson correlation	Kendall's Tau correlation is much more favourable than Spearman. Quadrant correlation demonstrates significantly inferior efficiency and even lower than Pearson
Tugran et al. (2015)	PCC is sensitive to outliers	PCC, Spearman, Kendall's Tau, Winsorized and Permutation-based correlation coefficient	Winsorized correlation coefficient and Permutation-based correlation are recommended to be better alternatives
Wilcox (2016)	PCC is not robust to outliers	PCC, Spearman, Winsorized correlation, Skipped correlation	Spearman outperformed others
Okwonu et al. (2020)	PCC highly vulnerable to outliers	PCC, Spearman, Weighted moment correlation coefficient (WMCC), Weighted mean based on Mahalanobis Distance, Modified weighted approach, Minimum covariance determinant (MCD)	Spearman can withstand up to 40% outliers compares to other. Mahalanobis approach is not robust.

In conclusion, overall, various studies yielded diverse results concerning the effectiveness of the correlation coefficients. Based on the summary provided, it can be noticed that Spearman's correlation was generally more robust and effective than PCC particularly in the presence of outliers. While PCC was commonly used, it is highly susceptible to outliers, which can significantly distort its results. Kendall's Tau was also highlighted as a favourable alternative, particularly in comparison to Spearman's correlation, although it may not be as robust as Spearman in all cases. Winsorized correlation was recommended as better alternatives to PCC due to their improved performance in handling outliers. Therefore, this study employed PCC,

Spearman, Kendall's Tau, and Winsorized correlation coefficients into the clustering algorithm, and their performance were being compared. The following section will discuss the linkage criteria, which is another pivotal step in the hierarchical clustering algorithm.

2.9 Linkage Method

In agglomerative hierarchical clustering method, the linkage method calculate the distance between groups of objects as a mechanism of the pairwise distances between the objects (Adams, Arseneau-Robar, Bonnell, Stead & Teichroeb, 2021). The distances between the new and old clusters can be calculated using several types of linkage method (D'Urso & Vitale, 2020) which were single linkage, complete linkage, average linkage and Ward's method. Single linkage joins clusters based on the distance between their nearest members while complete linkage occurs when groups are merged depend on the distance between their most distant members (Johnson & Wichern, 2019). As for the average linkage, clusters are joined or fused based on the average distance between pairs of members in the respective sets (Componation & Byrd, 2000).

2.9.1 Average Linkage

Average linkage was originally proposed by Sokal and Michener (1958) in order to address the problem of limitation of single and complete linkage. Single linkage was viewed as too liberal as it merged clusters when the two items from the clusters were similar with one another while complete linkage, contrastingly was considered too conservative as it needs complete links among all members of the groups to be combined (Blashfield & Aldenderfer, 1988). Thus, average linkage is intended to

signify a natural compromise between the single and complete linkage in order to postulate a better calculation of the distance amongst clusters.

Moreover, this linkage was also known as unweighted pair group method with arithmetic mean (UPGMA) and it determines the distance between all pairing objects on the basis of average distance (Yusniyanti et al., 2021). Tilevik (2017) highlighted that all distances were contributed equally to the average that was computed as this is an unweighted method.

2.9.2 Ward's Method

Ward's method which was originally presented by Ward (1963) proposed that the decision on selecting the clusters' pair to be merged at every phase should be based on the optimal value of an objective function. Ward (1963) stated that an objective function was any function that investigator chooses to reflect. In his paper, the objective function was the loss of information as reflected by the error sum of squares. The most desirable level of it was the minimum value. In addition, Ward's method or also known as Ward's minimum variance method (Szekely & Rizzo, 2005) took into account between-cluster and within-cluster distances. They stated that the within-cluster distances should be minimized as well as the between-cluster distances should be maximized. These two properties were combined into one criterion in Ward's method (Strauss & Von Maltitz, 2017).

2.9.3 Complete Linkage

Complete linkage method calculated the distance between clusters based on the greatest distance observed between two observations belonging to different clusters. This method has had a tendency to identify clusters that are compact and have equal

diameters (maximum distance between objects). However, it overlooks the consideration of the cluster structure (Everitt et al., 2011).

2.9.4 Single Linkage

Next, as for the single linkage, the distance between clusters was defined as the smallest distance between any observations belonging to different clusters. In large datasets, this method has had a tendency to generate unbalanced and scattered clusters, often referred to as "chained" clusters. This is because it does not take into account the structure of the cluster (Roux, 2018).

In addition, each linkage method has its own criteria of describing the distance. The chosen linkage will have a direct influence on the process of clustering and how the groups are combined. Consequently, the final cluster solution will be impacted (Yim & Ramdeen, 2015). A summary on the comparison of aforementioned linkages is shown in Table 2.3.

Table 2.3

Comparison of linkage methods

Method	Linkage criteria	Comment	Selected references
Single linkage	Distance between observations is the smallest	Chaining effect	Roux (2018), Yim and Ramdeen (2015)
Complete linkage	Distance between observations is the longest	Sensitive to noise	Emmendorfer and de Paula Canuto, (2021), Jarman (2020)
Average linkage	Average distance of all the observations	Best choice for most application	Emmendorfer and de Paula Canuto (2021), Jarman (2020), Tilevik (2017), Yim and Ramdeen (2015)
Ward's	The distance is how much	Give more accurate	Eszergar-Kiss and

method	the increase of sum of squares when two observations or clusters are being merged	cluster results	Caesar (2017)
--------	---	-----------------	---------------

Based on the Table 2.3, single linkage posed the problem of chaining effect while complete linkage was susceptible of outliers. This is due to their calculation of the distances of the observations between clusters whereby single linkage computed the minimum distance which yield unfavourable long clusters (Roux, 2018) and complete linkage calculated the maximum distance of all objects which produce compact clusters that was vulnerable to noises (D'Urso & Vitale, 2020; Emmendorfer & de Paula Canuto, 2021; Jarman, 2020). On the contrary, the average linkage method was said to be the best choice amongst other linkages while Ward's method falls second as it provided more accurate results of cluster. This is because, in the clustering process, average linkage computed average distance of all the objects (Yim & Ramdeen, 2015) and Ward's method took into account the variances in between clusters (Strauss & Von Maltitz, 2017).

2.9.5 Past studies on linkage methods

Comparing linkage methods using numerical methods has been a longstanding issue. Many past studies have been conducted to compare the performance of different linkage method in agglomerative hierarchical clustering. Beginning with the development of cophenetic correlations methods by Sokal and Rohlf (1962) for comparison of dendrograms have recently been the object of strong interest. Baker (1974) explored the influence of observational errors on dendrograms generated by complete linkage and single linkage hierarchical grouping techniques. The quality of fit for the dendrograms was assessed using the Goodman-Kruskal gamma coefficient.

The gamma coefficients revealed that the single linkage was more susceptible to the specific type of data errors employed compared to the complete linkage.

Blashfield (1976) compared four hierarchical clustering methods which are single linkage, complete linkage, average linkage, and Ward's method to determine their accuracy in recovering original population clusters by using Cohen's statistics. The results indicated that Ward's method outperformed the other clustering procedures significantly, while average linkage yielded relatively poor results. According to Milligan (1980) the introduction of outliers into the simulated data negatively affected the performance of complete linkage and Ward's method. Hands and Everitt (1987) studied the comparison between single linkage, complete linkage, average linkage, centroid linkage, and Ward's method on multivariate binary data. They discovered that Ward's method was the best among the other methods. Milligan and Cooper (1988) employed single linkage, complete linkage, group average (UPGMA) and Ward's method as a factor in their overall design to generate partition solutions. Their findings revealed that the single linkage was the least effective, while UPGMA and Ward's methods demonstrated the highest overall recovery rates.

Moreover, Ferreira and Hitchcock (2009) in their study had compared the performance of single linkage, complete linkage, average linkage, and Ward's method for clustering functional data by using Rand Index. The findings indicated that Ward's method generally outperformed the others, whereas average linkage showed superior performance in specific cases, particularly when the number of clusters exceeded the specified amount. Kuiper and Fisher (2011) in their research conducted a comparison of six hierarchical clustering procedures which were single linkage, complete linkage, median, average linkage, centroid, and Ward's method for

multivariate normal data. They made the assumption that the true number of clusters was known. To evaluate the clustering methods, they utilized the Rand index, which measures the proportion of correct groupings. In their investigation, for clusters of equal sizes, Ward's method and complete linkage method were found to be the most effective, while centroid and average linkage method performed best in the case of very unequal cluster sizes. The summary of the literatures is formulated as in the Table 2.4.

Table 2.4

Summary of linkage method literatures

Authors	Linkage methods being studies	Validity measures used	Findings
Baker (1974)	Complete linkage and single linkage	Goodman-Kruskal gamma coefficient	Single linkage displayed a higher vulnerability to the particular type of data errors used compares to complete linkage
Blashfield (1976)	Single linkage, complete linkage, average linkage, and Ward's method	Cohen's statistic	Ward's method outperformed others while average linkage produced quite poor results.
Milligan (1980)	Single linkage, complete linkage, Group Average, Weighted Average, Centroid method, Median method, Ward's method, Beta-Flexible Average Link in the New Cluster, Minimum Total Sum of Squares, Minimum Average Sum of Squares, MacQueen's method, Forgy's method, Jancey's method,	Rand index	Complete linkage and Ward's method are negatively affected by outliers. K-means perform excellently in all types of conditions

	Convergent K-means		
Hands and Everitt (1987)	Single linkage, complete linkage, average linkage, centroid, and Ward's method	EM algorithm	Ward's method was the best among others
Milligan and Cooper (1988)	Single linkage, complete linkage, group average (UPGMA) and Ward's method	Rand Index	Single linkage displayed the lowest effectiveness. UPGMA and Ward's methods exhibited the highest overall rates of recovery.
Ferreira and Hitchcock, (2009)	Single linkage, complete linkage, average linkage, and Ward's method	Rand Index	Ward's method performed better than others. Average linkage performed well specifically when the number of clusters surpassed the stipulated amount.
Kuiper and Fisher (2011)	Single linkage, complete linkage, median, average linkage, centroid, and Ward's method	Rand index	Ward's method and complete linkage are effective for cluster of equal sizes. Centroid and average linkage exhibited optimal performance when dealing with unequal cluster sizes.

Based on the Table 2.4, it can be observed that Ward's method consistently emerged as the most effective linkage method across various studies and validity measures. It is frequently noted for its superior performance in recovering true cluster structures and generally outperforms other methods, particularly in diverse clustering scenarios. Complete linkage also showed effectiveness in some cases, especially when dealing with clusters of equal sizes. However, it is susceptible to outliers, as indicated in some studies. Single linkage is consistently identified as the least effective, particularly vulnerable to data errors and generally underperforming compared to other methods. Average linkage shows mixed results, performing poorly in some studies but relatively well when the number of clusters exceeds a certain amount or when clusters are of unequal sizes. Therefore, in this study, all of these four linkage

methods had been used and paired with four correlation-based distances in order to compare their performance across few data conditions.

In conclusion, as a whole, one of the issues in clustering is, there is no ideal algorithm that can be applied to produce the ultimate data segmentation (Jarman, 2020). There are a lot of clustering techniques have been introduced since the cluster cannot precisely define the data. Usually, it depends on the perspective of researchers (Estivill-Castro, 2002). This limitation has been faced by previous researchers because there was no measurement to indicate the best clustering algorithm. Therefore, given the existing studies in the literature and the significance of selecting the most appropriate cluster method based on various conditions such as sample size, number of variables, and distance measures, a comprehensive simulation experiment was conducted. This study offers deeper understanding of how the cluster method operates in diverse conditions. The objective of this research was to explore the appropriate clustering method across different data conditions.

3.0 Water Quality assessment using cluster analysis

Water quality assessment is crucial for ensuring the sustainability of water resources. Cluster analysis, a multivariate statistical technique, has been widely employed in past studies to categorize water quality data into meaningful clusters. For example, Singh, Malik, Mohan and Sinha (2004) had conducted a study whereby in their assessment of Gomti river in India, cluster analysis, factor analysis/principal component analysis and discriminant analysis was being compared and used to group sampling sites based on water quality parameters. The study considered 24 parameters like Dissolved Oxygen (DO), chemical oxygen demand (COD) and others. Their findings revealed that cluster analysis had helped to cluster the eight

sampling sites into three clusters with same characteristics pertaining to water-quality parameters. They asserted that cluster analysis was an effective technique in providing reliable classification of surface waters across the entire region, enabling the design of an optimal future spatial sampling strategy. Consequently, the number of sampling sites in the monitoring network was reduced, thereby lowering costs without losing much information.

Panda, Sundaray, Rath, Nayak and Bhatta (2006) had used factor and cluster analysis to assess the water quality of the Mahanadi River, India and its estuarine systems. The findings showed that dendrogram produced by cluster analysis had helped to categorize different level of contamination and the effect of contamination into the stations while factor analysis revealed that anthropogenic contribution of nutrients was responsible for lowering some parameter such DO and pH levels in the water. They also had emphasized that cluster analysis technique was effective in providing reliable classification of both fresh and estuarine water systems across the entire region. This method facilitates the design of an optimal future spatial sampling strategy, ensuring accurate and conclusive results.

Moreover, Al-badaii, Shuhaimi-Othman and Gasim (2013) had carried out a study to determine the Semenyih River water quality in Selangor, Malaysia based on the physicochemical and biological parameters such as DO, COD, pH level, temperature and more by using hierarchical cluster analysis. The results showed that cluster analysis had helped to classified 8 sampling stations into three clusters based on the similarities of water quality features. They highlighted that this finding can help the authorities and decision makers to have the latest information on the river, as it

assists them in establishing optimal strategies that can reduce the number of sampling stations and experimental analysis costs.

The study by Masthurah, Juahir and Zanuri (2021) had investigated the spatial water quality assessment of Juru River Penang Basin, Kuantan River Basin and Johor River Basin in Malaysia by using cluster analysis, principal component analysis and discriminant analysis. 13 water quality parameters had been selected and analysed by using those methods in order to determine the origin of pollution sources on water quality data of those three river basins. Based on the findings, they stated that cluster analysis had successfully classified three clusters that designated 3 levels of pollution source, high, medium and low in each of the three river basins. These classifications enable the development of sampling strategy that can reduce the number of sampling stations and monitoring costs as one station per cluster is sufficient to accurately represent the spatial water quality for the entire region within that group.

In summary, based on the past literatures given, it is evident that cluster analysis is a valuable tool and had been massively applied in the contexts of water quality assessment, providing insights into pollution patterns and aiding in effective environmental management. Therefore, in this study, we utilized water quality data from the Juru River Basin for practical application, as previous research has demonstrated that cluster analysis was effective in assessing water quality. Besides, Juru River Basin in Pulau Pinang was deemed to be among the polluted rivers in Malaysia in the year 2010 (Zainudin, 2010) and 2013 (Mei, Chuan & Ling, 2016). By applying cluster analysis to this real-world dataset, we aimed to validate its utility in identifying patterns and variations in water quality across the basin.

3.1 Clustering validation measures

After performing the clustering technique, it is essential to validate the clusters result as the number of clusters being produced do not yet have a solid basis for determining the optimal number of clusters. Cluster validation involves quantitatively evaluating the outcomes of the clustering technique to identify the most suitable or optimal number of cluster (Fernandes, 2021). An ideal cluster was characterized by closely spaced individuals within the group while maintaining a clear separation from other clusters (Patibandla & Veeranjanyulu, 2018).

There are two types of cluster validation measures which are external clustering validation and internal clustering validation (Dubes, 1987). The primary difference between this two lies in the utilization of external information for validating the cluster. In contrast to external validation measures that rely on external information not inherent in the data, internal validation measures solely depend on the information contained within the data (Halkidi, Batistakis & Vazirgiannis, 2002). In simple words, internal measures assess the quality of a clustering structure without considering external information. In addition, external validation measures are primarily employed to select an optimal clustering algorithm for a given dataset as they have prior knowledge of the “true” number of clusters. Conversely, internal validation measures can be utilized to determine both the best clustering algorithm and the optimal number of clusters without any additional information (Liu, Li, Xiong, Gao & Wu, 2010). Thus, in this study, we were focusing more on the internal validation measures.

In general, there are numbers of internal clustering validation measures such as Davies-Bouldin index (Davies & Bouldin, 1979), Dunn Index (Dunn, 1974) , Rand

Index (Rand, 1971), Folkes and Mallow index (Fowlkes & Mallows, 1983), Cophenetic correlation coefficient (Sokal & Rohlf, 1962), Silhouette index (Rousseeuw, 1987), Hartigan index (Hartigan, 1973), Correction to Hartigan index (Albalade & Suendermann, 2009) and Calinski-Harabasz index (Calinski & Harabasz, 1974).

On the whole, the primary focus index in this study is Cophenetic Correlation Coefficient (CCC). In using CCC, the cophenetic matrix, P_C , containing cophenetic distances was used to represent the hierarchical diagram produced by hierarchical algorithm. Thus, throughout CCC, a statistical index to compute the level of similarity between cophenetic matrix, P_C and P , the proximity matrix that contains similarities was established (Halkidi, Batistakis & Vazirgiannis, 2001). Cophenetic distances are the distances between all pairs of data points as measured in the dendrogram. Essentially, for each pair of data points, there is the height at which their individual clusters were merged in the dendrogram (Palacio-Nino & Berzal, 2019). Next, CCC ranges between -1 and +1, where a higher value indicates better preservation of distances. Since its introduced by Sokal and Rohlf (1962), CCC has been extensively applied in many studies using real and simulated dataset over the years. It served as a measure to evaluate the goodness of fit of the dendrograms generated, the degree of fit for classifying set of data and also as a criterion to assess the effectiveness of different clustering techniques (Carvalho et al., 2019; Farris, 1969). For instances, Rincon, Johnson, Crossa and Taba (1996) and, Dubreuil and Charcosset (1998) had used CCC to evaluate few clustering methods in grouping maize accessions on the basis of agronomic. Erisoglu and Sakallioğlu (2010) evaluated several distance measures on hierarchical clustering algorithms using CCC and Wilks' lambda test statistics via multivariate normal distribution simulated

datasets and real-life dataset retrieved from UCI Machine Learning Repository website. Saracli et al. (2013) assessed different linkage methods based on CCC using simulated data under two situations, with and without outliers. Next, Kumar and Toshniwal (2016) evaluated several agglomerative clustering algorithms using CCC on road accident dataset. Tilevik (2017) evaluated various distance measures and linkage methods via CCC by using drug cytokine profiles. Dogan and Birant (2022) employed CCC to compare the performance of agglomerative hierarchical clustering and their proposed k-centroid link method using 24 different publicly available benchmark datasets. Thus, in this study, CCC has been used to evaluate the performance of several clustering algorithms constructed from different correlation-based distance measures and linkage method under different conditions of observations for simulated data and real water quality parameters sets of data.

Moreover, apart from CCC, another clustering validations indices namely, Silhouette index (S), Hartigan index (H), Correction to Hartigan index (Hc(K)) and Calinski & Harabasz index (CH) were used in validating the clustering model. Silhouette index, which was introduced by Rousseeuw (1987) assessed the clusters' quality based on the closeness of objects in a cluster with other objects in the nearest cluster (Amorim & Hennig, 2015; Wang & Xu, 2019). Silhouette value ranges from -1 to +1 where a value closer to +1 shows better quality of clusters. A value of 0 indicated that the data point was on the border between two clusters. A value of -1 indicated that the data point was misclassified, i.e., it belongs to the wrong cluster (Chiang & Mirkin, 2010). The definition of Silhouette Index is:

$$S_i = \frac{p(i) - q(i)}{\max \{p(i), q(i)\}} \quad (2.1)$$

where $p(i)$ is the average distance of the i^{th} object to all objects in the similar cluster; $q(i)$ is the smallest average distance of the i^{th} object to all objects in other clusters.

Next, the Hartigan Index was proposed by John A. Hartigan in 1973. It measured the quality of clustering by evaluating the optimal clustering number based on sum of squares as an index (Milligan & Cooper, 1985). Higher Hartigan index indicated better clustering quality. This index was useful for evaluating the effectiveness of clustering algorithms and comparing different clustering solutions. The index is described as follow:

$$H = (n - k - 1) (WSS/WSS_{k+1} - 1) \quad (2.2)$$

where n is the number of data points, k is the number of clusters and WSS is the within-cluster sum of squared distances. The term $(n - k - 1)$ is the penalty factor that is used to prevent an increase in monotonicity as the number of clusters increases.

After that, Albalade and Suendermann (2009) had introduced a correction to Hartigan index in order to enhance the alignment of the index. This correction index is

$$Hc(k) = H(k - 1) \quad (2.3)$$

It considered a modification of the optimum criterion by maximising $Hc(k)$. Put simply, this new criterion focused on maximizing the relative improvement at k with respect to $k - 1$, specifically in terms of decreasing dispersion. This enables the direct utilization of the corrected index $Hc(k)$, without the need to resort the

previous inversion of the index (Albalade, Suendermann & Minker, 2011). Thus, $H(k - 1) - H(k)$ index stands for the difference between the index of a clustering with k clusters and a clustering with $(k - 1)$ clusters. The bigger difference in the number of clusters signify the recommended number of clusters to be created or chosen.

As for Calinski & Harabasz index (CH), it was introduced by Calinski and Harabasz (1974) to evaluate the clusters' output based on the average between-cluster and within-cluster sum of squares. Like the Hartigan Index, the Calinski & Harabasz Index was useful for evaluating the effectiveness of clustering algorithms and comparing different clustering solutions (Lee et al., 2020). The index is defined as follows:

$$CH = \frac{\overline{BSS}/(k - 1)}{WSS/(n - k)} \quad (2.4)$$

where k is the number of clusters, n is the number of data points, $\overline{BSS}$ is the average between-clusters sum of squared distances and WSS is the sum of squared distances within the clusters. The Calinski & Harabasz Index ranges from 0 to infinity, with higher values indicating better clustering quality.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

Based on the issues highlighted in Chapter 1, it is concluded that there is no ideal correlation-based distance measures and linkage methods for forming agglomerative clustering, as each of them has their own strengths and weaknesses. For this reason, the performance of the agglomerative clustering algorithms is subjected to statistical evaluation. This chapter commenced with details of the implemented procedure, followed by description of the simulation study, and brief explanation on four (4) types of correlation-based distances, four (4) types of linkage methods and the cophenetic correlation coefficient used in simulation study. It then continues with illustrative example of the appropriate clustering model applied on the real dataset of Juru River basin water quality parameters. The final section of this chapter addresses the method used for validating the clustering model applied to the real data.

3.2 Procedure implemented

Correlation-based distance is a distance measure used in clustering to evaluate the similarity between data points based on their correlation coefficients (Xu et al., 2021). It is particularly useful in clustering observation where the features (or variables) are correlated (J. Chen et al., 2023).

The classical PCC, and three (3) robust correlation distances were implemented in constructing clustering models which were Spearman correlation, Kendall's Tau correlation, and Winsorized correlation coefficients. While, linkages methods applied were the average linkage, Ward's method, complete linkage, and single linkage. Both distances and linkage methods are two important parts in producing clustering

model. The general clustering algorithm is, let n is the data points or observations to be grouped into several number of clusters, then a distance matrix of size $n \times n$ will be produced before performing linkage clustering algorithm.

1. At the initial stage, each n observation is considered as a sole independent cluster.
2. Select **distance formula** for identifying the distance between two observations that produce $n \times n$ distance matrix
3. Select **linkage clustering algorithm** to merge between two difference clusters based on selected distance matrix.
4. Start from M clusters with each cluster containing one item at each iteration:
 - a) Using the cluster distance matrix, find the two nearest clusters.
 - b) Update the list of clusters by merging the two nearest clusters.
 - c) Update the matrix of clusters distances accordingly
5. Repeat until all the observations are formed in one cluster.

The results of the clustering model were graphically presented in the form of a dendrogram. Dendrogram as in Figure 3.1 is a two-dimensional diagram that represent the formation of hierarchical clustering which illustrated the hierarchical structure of the data, providing insights on the clusters' relationships (Boyko & Tkachyk, 2023). A dendrogram consists of connecting lines that link clusters together. At the bottom row of a dendrogram, there are leaves representing individual objects. Therefore, the number of leaf nodes in a dendrogram corresponds to the number of objects included in the cluster analysis. Typically, a cluster is defined as a combination of two or more leaves joined by a common node and a branch (Fahad et al., 2014).

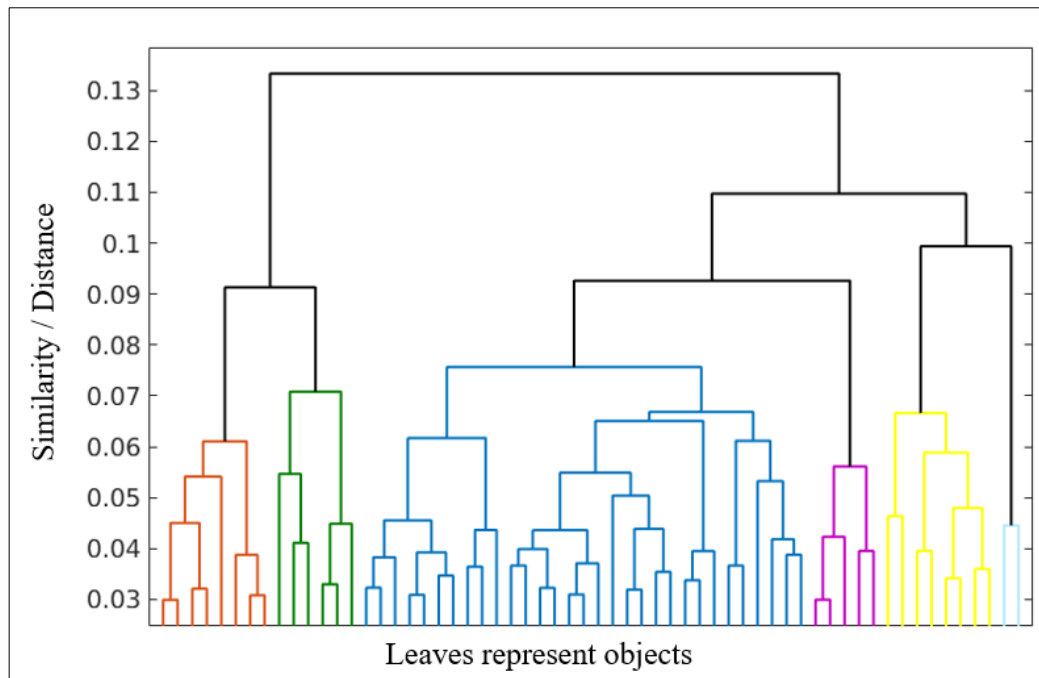


Figure 3.1 Dendrogram

In this thesis, simulation study and real data were used to answer the research questions. The statistical performance evaluation of sixteen (16) agglomerative hierarchical clustering algorithms were performed using simulation study to identify which appropriate algorithm is able to produce the good quality of clustering. The full research procedures are formulated in Table 3.1 as follows.

Table 3.1

Full detailed process of the research procedures

Simulation study	
Step 1	The data was generated based on these simulation conditions: i. Normal distribution ii. Normal distributions with mild and extreme contamination observations and different percentage of outliers iii. Non-normal distribution (Gamma distribution) iv. Different sample sizes v. Different number of variables
Step 2	Construct the clustering models based on four correlation-based distance measures and four linkage methods which produce 16 algorithms, respectively by using the simulation data conditions. The particulars are

as below:

Types of correlation-based distances used:

- i. PCC
- ii. Spearman correlation
- iii. Kendall's Tau correlation
- iv. Winsorized correlation

Types of linkage methods used:

- i. Average linkage
- ii. Ward's method
- iii. Complete linkage
- iv. Single linkage

16 clustering algorithms produced:

- i. PCC and average linkage
- ii. PCC and Ward's method
- iii. PCC and complete linkage
- iv. PCC and single linkage
- v. Spearman correlation and average linkage
- vi. Spearman correlation and Ward's method
- vii. Spearman correlation and complete linkage
- viii. Spearman correlation and single linkage
- ix. Kendall's Tau correlation and average linkage
- x. Kendall's Tau correlation and Ward's method
- xi. Kendall's Tau correlation and complete linkage
- xii. Kendall's Tau correlation and single linkage
- xiii. Winsorized correlation and average linkage
- xiv. Winsorized correlation and Ward's method
- xv. Winsorized correlation and complete linkage
- xvi. Winsorized correlation and single linkage

Step 3	Calculate average value of Cophenetic correlation coefficient (CCC) for each algorithm for 5000 iterations.
---------------	---

Step 4	Compare the statistical performance of the 16 clustering algorithms using CCC.
---------------	--

Step 5	Table of Summary was proposed
---------------	-------------------------------

Real data application

Step 6	Choose the appropriate clustering algorithm
---------------	---

Step 7	Construct the clustering model by applying the appropriate clustering algorithm chosen using Juru river basin water quality parameter dataset.
Step 8	Interpret the clustering model

The flowchart of the research procedures is shown in Figure 3.3

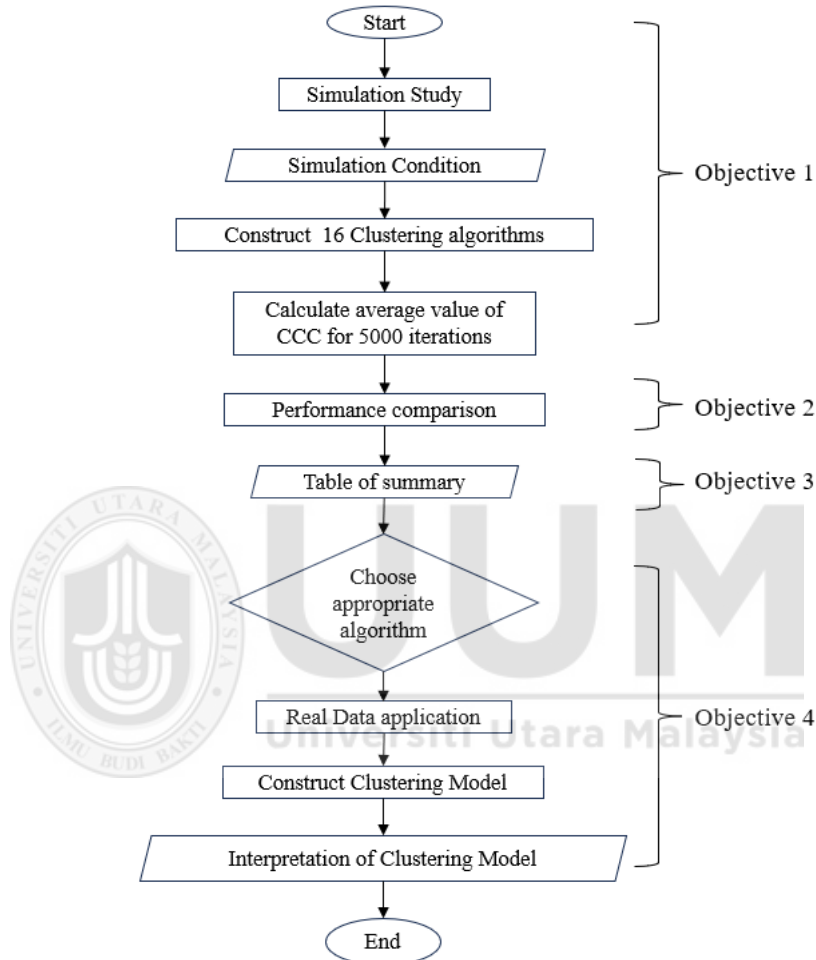


Figure 3.2 Flowchart of the full research procedures

The full research procedures which consisted two primary phases, the simulation study and the real data application were depicted in the flowchart in Figure 3.2. Based on the figure, the process started with the simulation study. Within this phase, the data were generated based on various simulation conditions. These conditions encompass both normal and non-normal data distributions, including the non-normal Gamma distribution. Additionally, several factors such as different sample sizes, the

number of variables, levels of data contamination, and the percentage of outliers were also being considered.

Subsequently, in order to achieve the first and second research objective, 16 clustering algorithms were employed to construct clusters based on these simulated conditions. To evaluate the performance of these algorithms, the average value of the CCC was calculated over 5000 iterations. In this study, 5000 iterations were being used in the context of simulation in order to ensure the stability and reliability of the results. Using 5000 iterations helps to ensure that the algorithm has sufficient time to converge, reducing the variability of the results and lead to more robust and reliable conclusions (Brooks & Gelman, 1998; Kim et al., 2000). Following the calculation of CCC values, a performance comparison was conducted among the 16 algorithms, focusing on their average CCC values. For the purpose of accomplishing the third research objective, the results of this comparison were then compiled and presented into a form of table of summary. This table also served as a basis for selecting the most appropriate clustering algorithm, a decision made in choosing the appropriate algorithm for the real data step.

The fourth research objective was attained in the next phase which involved the real data application. Here, the selected clustering algorithm was applied to a real dataset, specifically the Juru river basin water quality dataset. This application began with the construction of the clustering model using the chosen algorithm. The effectiveness and insights gained from this model were then assessed in the interpretation of the clustering model step, where the results were analysed and conclusions were drawn.

Finally, the process concluded at the end point, marking the completion of the research procedures. This structured and rigorous approach ensures that the selected

clustering algorithm was not only theoretically sound but also practically applicable to real-world data, providing meaningful insights and conclusions. In the next sections of this chapter, we discussed further details on how to conduct the simulation study for the 16 algorithms, and provide the illustrative applications on real data sets.

3.3 Simulation study

In statistical study, a simulation study refers to a research method in which a model is constructed to simulate data that is similar to real-world data (Dooley, 2002). After being developed or modified, the statistical technique is applied to the simulated data to draw conclusions about the underlying population for analysing the performance of the proposed technique (Morris, White & Crowther, 2019). Therefore, simulation study was used to investigate the behaviour of clustering models and its performance under different data conditions. Besides, we can investigate the impact of different distances, linkage methods and model specifications on clustering algorithm.

In this study, different data distributions, different sample sizes, different numbers of variables, different levels of data contamination, and percentage of outlier were considered to evaluate the quality of a clustering solution in hierarchical clustering. 5000 iterations has been used Based on 5000 iterations, the average value of CCC is calculated to measure how well the clustering solution preserves the pairwise distances between observations (Carvalho et al., 2019; Kumar & Toshniwal, 2016; Saracli et al., 2013).

The simulation conditions are as follows,

- i. Data distribution are normal distribution and non-normal distribution (skewed distribution).

- ii. Sample size, $n = 10, 50, 100, 200, 500$
- iii. Number of variables, $p = 3, 5, 10, 20, 30$
- iv. Three different conditions of distributions:
 - a. Normal distribution
 - b. Normal distributions with two different level of contaminated observations which are the control parameter value, $\mu = 3$ indicates mild contamination and $\mu = 5$ for extreme contamination (Alfaro & Ortega, 2009; Haddad, Syed-Yahaya & Alfaro, 2012). For each level of contamination, there are three different percentages of data contamination, 10%, 20%, 30%.
 - c. Non-normal distribution, i.e multivariate Gamma distribution. In this study the shape and scale parameters were set with $\alpha=3$ and scale $\beta=1$ (Khoo, Saha, Teh, Haq & Lee, 2020). Gamma distribution was being chosen in this study as it can effectively model skewed distributions where the data is not symmetrically distributed around the mean by adjusting the shape (α) and scale (β) parameters (Bhaumik, Kapur & Gibbons, 2009). Besides, Gamma distribution had been extensively used as it can represent a wide variety of shapes from nearly symmetric to highly skewed which make them a good fit and flexible in modelling different types of data (Chen, Buis & Zhao, 2023; Chen et al., 2014; Derya & Canan, 2012).

The following Table 3.2 list the general commands in MATLAB that were used to generate the data for three conditions of observations.

Table 3.2

List of general commands for simulated data in MATLAB

Conditions of observations	Command
Normal distribution $X = \text{mvnrnd}(\mu, S, n)$	<pre> r=5000; p=3; %p is number of variables n=10; %n is sample size mu=[0 0 0]; %mean vector S =[1 0 0; 0 1 0; 0 0 1]; %covariance </pre>
Normal distribution with mild and extreme contaminated observation $X = \text{mvnrnd}(\mu, S, n)$	<pre> %Mild contaminated observation r=5000; p=3; %p is number of variables n=10; %n -sample size - can change pi=0.3; %percentage of outlier mu =[3 3 3];%mean vector S=[1 0 0; 0 1 0; 0 0 1]; %covariance cd=pi*n; %no of contaminated data %Extreme contaminated observation r=5000; p=5; %p is number of variables n=10; %n -sample size - can change pi=0.3; %percentage of outlier mu =[5 5 5];%refer to number of variable sigma=[1 0 0; 0 1 0; 0 0 1]; %covariance cd=pi*n; %no of contaminated data </pre>
Non normal distribution- Gamma distribution $X = \text{gamrnd}(3, 1, [n \ p])$	<pre> r=5000; p=5; %p is number of variables n=10; %n is sample X = gamrnd(3,1,[n p]); %alpha=3, beta=1 </pre>

In this study, several simulation scenarios were designed to assess the performance of the algorithms under different conditions. These simulations were implemented in MATLAB using general commands to generate the required datasets. The simulation conditions were categorized into normal distribution, normal distribution with mild

and extreme contamination, and non-normal distribution, specifically using the Gamma distribution.

For the normal distribution condition, multivariate normal data were generated using $x = \text{mvnrnd}$. The parameters included a mean vector $\mu = [0 \ 0 \ 0]$ with the number of zeros corresponding to the number of variables; a covariance matrix of $S = [1 \ 0 \ 0; 0 \ 1 \ 0; 0 \ 0 \ 1]$ adjusted to match the number of variables, sample size, $n = 10, 50, 100, 200, 500$ and number of variables, $p = 3, 5, 10, 20, 30$. This was repeated $r=5000$ times.

For scenarios involving mild and extreme contamination, the mvnrnd function was again used. Mild contamination involved setting the mean vector $\mu = [3 \ 3 \ 3]$ and maintaining the same covariance matrix S (the setting of both μ and S were adjusted according to number of variables). A contamination proportion for example $\pi=0.3$ indicating 30% of outliers injected was applied to the sample size, n resulting in $cd=\pi*n$ contaminated data points. For extreme contamination, the parameters were adjusted to increase the severity. Here, the mean vector, μ was set for example $\mu = [5 \ 5 \ 5]$ with the same covariance matrix structure.

Lastly, for the non-normal distribution scenario, data were generated from a Gamma distribution using the gamrnd function. The shape parameter α was set to 3, and the scale parameter β was set to 1. Then, the sample size, n and the number of variables, p was set accordingly. This scenario was also repeated $r=5000$. Overall, the MATLAB commands provided a straightforward and efficient means of generating the diverse datasets required for this study. Next sections will provide further discussion on the type of correlation-based distances and linkage method used, CCC, real data application and CH index.

3.4 Correlation-based distances

In correlation-based distances, the similarity between clusters is calculated based on the similarity between their constituent data points (Xu et al., 2020). The correlation-based distances that are being used in producing clusters in this study were PCC, Spearman correlation, Kendall's Tau correlation and Winsorized correlation.

3.4.1 Pearson Correlation Coefficient (PCC)

PCC is the measure of association that is most commonly utilized (Puth, Neuhauser & Ruxton, 2014; Schober & Schwarte, 2018). It quantifies the strength and direction of the linear relationship between two variables. PCC is a parametric method that depends on several assumptions regarding the data which include the relationship between two variables should be linear, data is normally distributed, and no significant outliers in data (Armstrong, 2019; Gogtay & Thatte, 2017; Kozak et al., 2012; Li & Kim, 2003; Schober & Schwarte, 2018). Kozak et al. (2012) and Armstrong (2019) stated that when the relationship between the variables is non-linear, the correlation coefficient might not accurately reflect the strength and direction of the association. In addition, if the data is not normally distributed, it is preferable to calculate correlation coefficients based on the ranks of the data rather than their actual values, whereby non-parametric methods like Spearman correlation should be used (Akoglu, 2018; Alsaqr, 2021; Armstrong, 2019). Additionally, outliers can impact the performance of PCC, leading to potential misinterpretations of the coefficient and inaccuracies of analysis (Gogtay & Thatte, 2017; Kim et al., 2015; Kozak et al., 2012; Mukaka, 2012; Schober & Schwarte, 2018). Apparently, these assumptions were often violated in real situation and to obtain the ideal data which satisfy all the assumptions is hardly achieved (Wu, 2007). Thus, few approach and alternatives such as non-parametric and robust method had been proposed.

Moreover, PCC values range from -1 to +1 which -1 denotes a perfect negative linear relationship and +1 denotes a perfect positive linear relationship between variables. There is no linear relationship between the variables when PCC is 0. Let X and Y be random variables where $x_i, y_i, i = 1, 2, \dots, n$ are the observations from those two variables. The population PCC, ρ is defined as follows,

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (3.1)$$

where the population correlation is lies between -1 and 1, $-1 < \rho < 1$, σ_{XY} is the covariance between two random variables X and Y , σ_X is the standard deviation of X and σ_Y is the standard deviation of Y (Wiedermann & Hagmann, 2016). As for the sample PCC, r , the equation is as given below,

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2]^{1/2}} \quad (3.2)$$

where n is the sample size, x_i and y_i are the i -th sample observations and $\bar{x}$ and $\bar{y}$ are the sample means for the random variables X and Y , respectively (Tilevik, 2017).

3.4.2 Spearman correlation

Spearman's correlation coefficient is a measure of the strength and direction of the monotonic relationship between two variables. It ranges from -1 to +1. A value of -1 indicates a perfect negative monotonic relationship, 0 indicates no monotonic relationship, and +1 indicates a perfect positive monotonic relationship. Monotonic relationship refers to the situation where, as one variable increases, the other variable either also increases or decreases, but not necessarily at a constant rate (El-Hashash & Shiekh, 2022; Schober & Schwarte, 2018).

Let $R(X_i)$ and $R(Y_i)$ denote the ranks of a pair of observations X and Y $((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n))$ where each rank is integers. Each X and Y were ranked relative to all other observed values of X and Y , from high to low or vice versa based on their magnitudes (Zar, 2005). According to Zar (2005) and Tugran et al. (2015), this non-parametric correlation coefficient is computed as,

$$r_{XY} = \frac{\sum_{i=1}^n R(X_i)R(Y_i) - n(\frac{n+1}{2})^2}{\sqrt{(\sum_{i=1}^n R(X_i)^2 - n(\frac{n+1}{2})^2)(\sum_{i=1}^n R(Y_i)^2 - n(\frac{n+1}{2})^2)}} \quad (3.3)$$

Alternatively, it can be expressed as

$$r_{SXY} = 1 - \frac{6 \sum_{i=1}^n (d_i)^2}{n(n^2 - 1)} \quad (3.4)$$

where n is the number of samples and d_i is the difference for each pair of ranks. When the rank of X , $R(X)$ is the same as the rank of Y , $R(Y_i)$ for every pair of observations, d_i will be equal to zero.

3.4.3 Kendall's Tau correlation

Kendall's Tau correlation (Kendall, 1938) is a correlation coefficient that measures the strength and direction of association between at least two ordinal variables. It ranges from -1 to +1, where a value of -1 indicates a perfect negative association, 0 indicates no association, and +1 indicates a perfect positive association (Temizhan, Mirtagioglu & Mendes, 2022)

Consider $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ represent a collection of combined observations derived from two random variables X and Y . For any pair of observations $(X_i, Y_i), (X_j, Y_j)$, if both $X_i > X_j$ and $Y_i > Y_j$ or if both $X_i < X_j$ and $Y_i < Y_j$

Y_j , they are known to be concordant as the ranks for both pairs agree. Meanwhile, they are call discordant if $X_i > X_j$ and $Y_i < Y_j$ or if $X_i < X_j$ and $Y_i > Y_j$. Thus, the formula of Kendall's Tau correlation coefficient is as follows (Rajesh & Chaturvedi, 2019; Tugran et al., 2015),

$$rk_{XY} = \frac{n_c - n_d}{n(n-1)/2} \quad (3.5)$$

where n_c and n_d is the number of concordant and discordant pairs while n is the number of pairs of observations. To obtain n_c and n_d , first, arranged the observations (X_i, Y_i) so that X are in natural (ascending) order and numbers of Y pairs in natural and reverse natural order with respect to each Y . Thus, n_c will be total value of Y in natural order and n_d is the total value if Y in reverse natural order. Thus, if $n_c > n_d$, then there is a direct positive correlation between the X and Y rankings. In contrast, if $n_c < n_d$, then there is a negative inverse correlation between X and Y rankings.

3.4.4 Winsorized correlation

Another robust analogue of PCC is the Winsorized correlation (Wilcox, 1993) that can be computed based on two random variables X_1 and X_2 . Like the Pearson correlation, Winsorized correlation ranges from -1 to 1. A positive value indicates a positive relationship, while a negative value indicates a negative relationship between the two variables (Wilcox, 1994). The computation of Winsorized correlation uses PCC's formula after Winsorizing the data (Wilcox, 2016). To compute Winsorized correlation based on the random sample $(X_{11}, X_{12}), \dots, (X_{n1}, X_{n2})$, first winsorize the observations by computing the Y_{ij} values (Alkenani & Yu, 2013; Wilcox, 2016) as presented below:

$$Y_{ij} = \begin{cases} X_{(g+1)j} & \text{if } X_{ij} \leq X_{(g+1)j} \\ X_{ij} & \text{if } X_{(g+1)j} < X_{ij} < X_{(n-g)j} \\ X_{(n-g)j} & \text{if } X_{ij} \geq X_{(n-g)j} \end{cases} \quad (3.6)$$

where g is the number of observations that are winsorized from each end of the distribution corresponding to the j th group. Then, the sample Winsorized correlation is as follows,

$$r_w = \frac{\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{[\sum_{i=1}^n (Y_{i1} - \bar{Y}_1)^2 \sum_{i=1}^n (Y_{i2} - \bar{Y}_2)^2]^{1/2}} \quad (3.7)$$

where n is the sample size, Y_{i1} and Y_{i2} are the i -th sample observations and $\bar{Y}_1$ and $\bar{Y}_2$ are the sample winsorized mean for the random variables Y_{i1} and Y_{i2} respectively.

3.5 Linkage method

The next important steps in agglomerative clustering is to choose the linkage method in order to merge potential clusters based on their selected specific metric (Tokuda, Comin & Costa, 2020). In this study, the linkage methods that were being studied are average linkage, Ward's method, complete linkage, and single linkage.

3.5.1 Average linkage

Average linkage involves computing the average distance between each pair of observations in different clusters and then merging the two clusters with the smallest average distance (Vijaya, Sharma & Batra, 2019). The computational distances between cluster AB and another cluster C by using average linkage is as follows,

$$d_{(AB)C} = \frac{\sum_i \sum_k d_{ik}}{N_{(AB)}N_C} \quad (3.8)$$

where d_{ik} is the distance between object i and object k in the cluster AB and cluster C , respectively. While $N_{(AB)}$ and N_C are the objects size in cluster AB and C , respectively.

3.5.2 Ward's Method

Ward's method was introduced by Ward (1963) involves minimizing the total within-cluster between all observations within each cluster. It combines the two clusters A and B that minimize the increase in the sum of squared errors (Bu, Liu, Pan & Ling, 2020). Assume that number of samples, n are categorized into k groups, the i –th sample in the cluster G_A is denoted as $X_i^{(A)}$, and n_A represents the number of samples in G_A . Let the centroid of the cluster be $\bar{X}^{(A)}$. Then, the sum of squares within G_A is:

$$S_A = \sum_{i=1}^{n_A} (X_i^{(A)} - \bar{X}^{(A)})' (X_i^{(A)} - \bar{X}^{(A)}) \quad (3.9)$$

The total sum of squares of k groups is:

$$S = \sum_{A=1}^k S_A = \sum_{A=1}^k \sum_{i=1}^{n_A} (X_i^{(A)} - \bar{X}^{(A)})' (X_i^{(A)} - \bar{X}^{(A)}) \quad (3.10)$$

In this method, n samples are initially considered as separate clusters in which each cluster contains single number of samples. When each time two clusters combined, the number of clusters decreases by one and total sum of squares, S increases. This merging process aims to minimize the increase in S , which is a measure of the variance within the clusters. This is because the goal of Ward's method is to achieve a hierarchical clustering that minimizes the overall within-cluster variance, thereby

producing clusters that are as homogeneous as possible (Murtagh & Legendre, 2014; Strauss & Von Maltitz, 2017).

Thus, the distance between G_A and G_B is defined as the sum of squares between the two clusters:

$$D^2(G_A, G_B) = S_C - S_A - S_B \quad (3.11)$$

where $G_C = \{G_A, G_B\}$. The distance between the new cluster G_C and another cluster $G_k (k \neq A, B)$ is calculated as:

$$D^2(G_C, G_k) = \frac{n_k + n_A}{n_C + n_k} D_{kA}^2 + \frac{n_k + n_B}{n_C + n_k} D_{kB}^2 - \frac{n_k}{n_C + n_k} D_{AB}^2 \quad (3.12)$$

where:

n_k = number of samples in k cluster

n_A = number of samples in A cluster

n_B = number of samples in B cluster

n_C = number of samples in C cluster

n_r = number of samples in r cluster

3.5.3 Complete Linkage

Complete linkage involves calculating the maximum distance between each pair of observations in different clusters and then merging the two clusters with the smallest maximum distance (Mamun, Aseltine & Rajasekaran, 2016). The maximum distance between cluster AB and another cluster C are calculated by

$$d_{(AB)C} = \max\{d_{AC}, d_{BC}\} \quad (3.13)$$

Here, d_{AC} and d_{BC} are the maximum distances between observations of clusters A and C and clusters B and C respectively.

3.5.4 Single Linkage

Single linkage clustering calculates the minimum distance between each pair of observations in different clusters and then merging the two clusters with the smallest minimum distance (Ros & Guillaume, 2019). The distance is computed as

$$d_{(AB)C} = \min \{d_{AC}, d_{BC}\} \quad (3.14)$$

whereby d_{AC} and d_{BC} are the distances between the closest observations of clusters A and C and clusters B and C respectively.

3.6 Cophenetic Correlation Coefficient (CCC)

The cophenetic coefficient correlation was introduced by Sokal and Rohlf (1962). It is a measure of how well a dendrogram (tree diagram) represents the pairwise distances between objects or observations. It is calculated by correlating the pairwise distances between all objects in a dataset with the distances between the objects in the dendrogram (Palacio-Nino & Berzal, 2019).

The cophenetic correlation coefficient can range from -1 to +1, with values closer to +1 indicating a better fit between the dendrogram and the pairwise distances (Saracli & Aksit, 2022). A value of 1 indicates a perfect fit, while values closer to 0 or negative indicate poor fits. In addition, according to Rohlf (1970) and Carvalho et al. (2019), in practice, CCC value that are less than 0.7 would imply that the grouping method is inadequate for summarizing the information within the dataset. The computation of CCC is as given by:

$$CCC = \frac{\sum_{i < j} (U_{ij} - \bar{u})(V_{ij} - \bar{v})}{\sqrt{\sum_{i < j} (U_{ij} - \bar{u})^2} \sqrt{\sum_{i < j} (V_{ij} - \bar{v})^2}} \quad (3.15)$$

where U_{ij} is the distance between observations i and j in U , $\bar{u}$ is the average of U_{ij} , V_{ij} is the cophenetic distance or dendrogram distance between observations i and j and $\bar{v}$ is the average of V_{ij} . This distance is the height of the node at which these two points are first joined together.

3.7 Real Data Application on The Sample of Juru River Basin Water Quality.

Water quality is a critical aspect of environmental monitoring and public health. Using real-world water quality data ensures that the clustering algorithm is tested on data that has practical relevance, making the findings directly applicable to real-world scenarios (Du et al., 2017). In addition, water quality datasets typically include multiple parameters such as pH, dissolved oxygen, chemical oxygen demand, biological oxygen demand, nitrate levels, and more. This high-dimensional data is ideal for testing clustering algorithms, which are designed to handle and identify patterns in multivariate data (Shrestha & Kazama, 2007).

Therefore, in this study, the real dataset of water quality parameters is used to illustrate the application of the suitable clustering model. The bimonthly data of five water quality parameters of Juru River Basin for 57 observations which are Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Suspended Solids (TSS), and Ammoniacal Nitrogen (NH₃-NL) were collected for the past 10 years, from 2008 until 2017 from the Department of Environment Malaysia (DOE). These five parameters were being chosen because they are the parameters that are primarily used in national's Water Quality Index (Ho et al., 2019). Water Quality Index (WQI) assigns quality values to a collection of measured parameters by utilizing sub-index values. These sub-index values are allocated to pre-identified parameters based on a comparison of their measurements

with parameter-specific rating curves. Optionally, the values can be weighted and combined to form the final index (Zainudin, 2010). Besides, WQI also serves as a highly effective tool for offering feedback to policymakers and environmentalists regarding water quality. It assesses the overall water quality status at specific times and locations, providing valuable information (Naubi, Zardari, Shirazi, Ibrahim & Baloo, 2016).

In addition, as for the water quality parameter, DO quantifies the level of freely available oxygen in water, typically measured in milligrams per litre (Naubi et al., 2016). It is influenced by temperature, with colder water having a greater capacity to hold oxygen. Insufficient DO levels in river water cause aquatic species to disperse, weaken, or potentially perish. Next, BOD also serves as the primary indicator of pollution resulting from the disposal of domestic and industrial effluents, organic sources, and sewage treatment, which often contain significantly elevated levels of bacteria (Abdul-Karim & Sharif, 2019). Moreover, it is presumed that the presence of COD in the river basins originates from anthropogenic activities associated with nearby industrial areas, where the discharge of industrial waste into the river occurs (Masturah, Juahir & Mohd Zanuri, 2021). TSS are naturally occurring pollutants that contribute to the turbidity of river water. Additionally, an excessive presence of TSS in water can indicate land erosion within the river catchment. Lastly, NH₃-NL parameter is employed to identify contaminants that stem from the practice of agricultural fertilisers, alongside animal waste and household sewage (Hashim et al., 2019). The deterioration of water quality caused by NH₃-NL remains a significant global environmental and public concern due to its potential to trigger eutrophication.

Juru River Basin in Pulau Pinang is among the polluted rivers in Malaysia in the year 2010 (Zainudin, 2010) and 2013 (Mei, Chuan & Ling, 2016). This river has an area of 75km^2 and comprises of four rivers which are Juru River, Kilang Ubi River, Pasir River and Rambai River that streams into the Straits of Malacca (Abdul Zali, Retnam & Juahir, 2011). The water of Juru River Basin was polluted because of numerous activities such as agricultural and industrial activities, wastes produced from domestic activities, and development of land activities (Karim, Sharif, Zakaria & Azis, 2019; Mei-Wo & Mahmood, 2020). Subsequently, the deterioration of the water quality of this river will disrupts the ecosystem by damaging the habitat of streams species and can also lead to the crisis of water supply (Masthurah et al., 2021).

The steps of application on real data sets were shown in Figure 3.3.

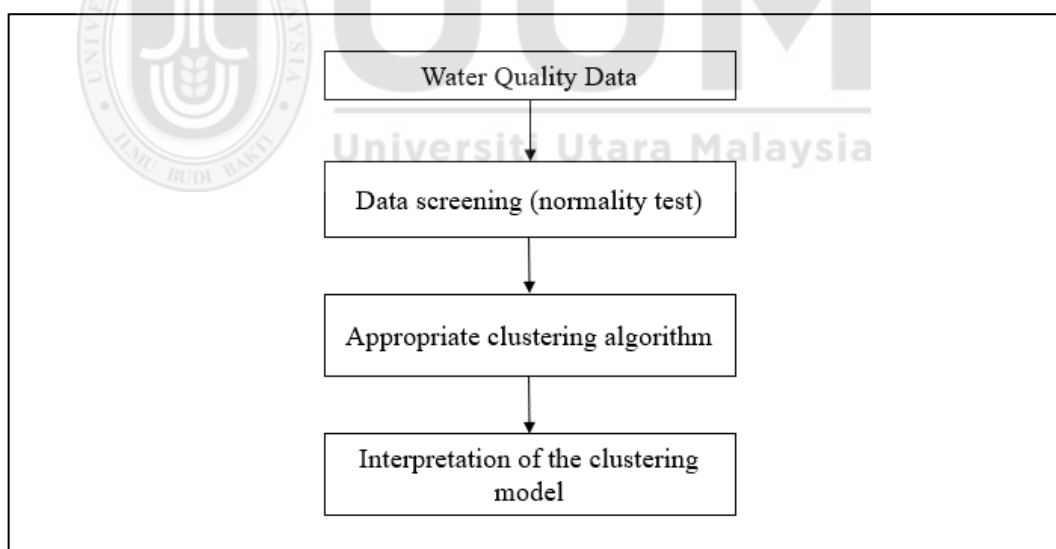


Figure 3.3 Steps of application on real data

Based on Figure 3.3, the process began with the collection of water quality data which consist of five water quality parameters which are DO, BOD, COD, TSS, and NH₃-NL from the website of Department of Environment Malaysia (DOE). Then, before applying the clustering algorithm, data screening, which included conducting

a normality test, was performed on the real data to check whether the data follows a normal distribution assumption. If the data is not normally distributed, appropriate clustering algorithm that suits the data will be applied. In this study, the results of data screening showed that the water quality data were normally distributed. After screening the data, the appropriate clustering algorithm was selected based on the characteristics of the data. Hence, based on the simulation study, the appropriate clustering algorithm under normal distribution when $p = 5$ and $n = 50$ was *Spearman-average* had been chosen to be implemented on real data. This is because the real data also consists the same $p = 5$ and the $n = 57$. The final step is the interpretation of the clustering model. This involves evaluating the cluster performance based on CCC and CH index and analysing the results which involved examining the spider web, correlation matrix, evolution of indices, results by cluster, dendrogram plot, heat map analysis and cluster profile plot. This step was done in order to understand the underlying patterns and relationships within the water quality data.

3.8 Calinski & Harabasz (CH) index

Lastly, the clustering model is validated by using internal validation measures. Calinski & Harabasz (CH) index was used to evaluate the effectiveness of a clustering solution. It was proposed by Calinski and Harabasz (1974) as a way to assess the "goodness" of a clustering solution based on the between-cluster sum of squares (BSS) or also known as inter-cluster divergence/covariance and the within-cluster sum of squares (WSS) or intra-cluster covariance. The higher the BSS, the higher the degree of dispersion between the clusters is while the smaller WSS results in closer relationship within the cluster (Wang & Xu, 2019). The index is defined as in equation 2.4 in previous Chapter 2.

CHAPTER FOUR

ANALYSIS AND FINDINGS

4.1 Introduction

In this chapter, the results of the simulation study and real data application were presented. This chapter reported on the results of statistical performances of 16 clustering algorithms based on average value of CCC. From simulation study, the appropriate clustering algorithm was chosen to illustrate the real data application using Juru river basin water quality dataset. The analysis on clustering model includes spider web, correlation matrix, evolution of indices, results by cluster, heatmap analysis, and cluster profile plot. All the results were analysed by using MATLAB R2023a and XSLTAT software.

4.2 Results of simulation study

This section displayed and analyzed the statistical performances based on the average value of CCC for the 16 clustering algorithms under three different conditions of distributions. Based on the statistical performance, a series of analysis was run under simulation program for the 16 algorithms formed from the combination of four different correlation-based distances, namely PCC, Spearman correlation, Kendall's Tau correlation, and Winsorized correlation with four linkage methods which were average linkage, Ward's method, complete and single linkage. In the simulation program, the simulation condition involved three different data distributions which were normal distribution, normal distribution with mild and extreme level of data contamination injected with 10%, 20% and 30% percentage of outlier for each level, and non-normal distribution which was multivariate Gamma distribution. The sample sizes used was $n = 10, 50, 100, 200$ and 500 while the number of variables, $p = 3, 5,$

10, 20 and 30. Then, their performances were evaluated based on the average value of CCC.

Cophenetic Correlation Coefficient (CCC) is being employed in this study to assess the degree to which the dendrogram accurately reflects the similarities or dissimilarities among the observations that have been clustered. CCC value that is closer to 1 shows the clustering result is almost accurate. As stated by Rohlf (1970) in practice, if the CCC value falls below 0.7, it indicates that the grouping method is defective to summarize the information contained in the dataset. Additionally, the 'rule of thumb' for interpretation of the strength of the correlation coefficient is employed as the guideline in the analysis of the CCC results. This is because the rule of thumb, as illustrated in Table 4.1, provides a clearer and more specific understanding of the direction and strength of the relationship between the variables (Schober & Schwarte, 2018).

Table 4.1

Rule of Thumb for the interpretation of correlation coefficient

Size of Correlation	Interpretation	References
0.90 to 1.00	Very high correlation	
0.70 to 0.89	High correlation	Asuero, Sayago and
0.50 to 0.69	Moderate correlation	Gonzalez (2006), Mukaka,
0.30 to 0.49	Low correlation	(2012), Schober and
0.00 to 0.29	Negligible correlation	Schwarte (2018)

Next, the numerical results that are gained from the simulation procedure are presented in the following three subtopics which are clustering algorithm under normal distribution, clustering algorithm under mild and extreme contaminated distribution, and clustering algorithm under multivariate Gamma distribution. In this

result analysis, clustering algorithm that has average CCC value that was 0.70 and above was considered to be as dominant algorithm, or having large coefficient, high correlation and also well-performed. This is signified with the bold-black font value in the table. The average CCC value that was between 0.3 to 0.699 is considered to be moderate correlation and no bold is applied on the black font value. Meanwhile, when the average CCC value was below 0.3, the clustering algorithm is considered to has low correlation and not well-performed. This will be indicated by the value in blue colour font. In addition, the largest coefficient for each sample size will be marked with star (*) on the left side of the average CCC value. As for the selection for the appropriate clustering throughout the table, the value that had or most had (*) will be chosen as the appropriate clustering algorithm indicating it had the highest coefficient value or had more frequency of highest coefficient of average CCC value.

4.2.1 Clustering Algorithm Under Normal Distribution

In this section, the simulation results under normal distribution is presented. The analysis has been done for $p = 3, 5, 10, 20, 30$ and $n = 10, 50, 100, 200, 500$. The statistical performance of clustering was presented in the following tables.

Table 4.2

Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 3$.

Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient (CCC)					
		$p = 3$					Freq
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.854	0.771	0.726	0.746	0.727	5
	Ward's	0.894 (*)	0.749	0.721	0.764	0.711	5
	Complete	0.845	0.775	0.679	0.686	0.652	2
	Single	0.763	0.691	0.618	0.633	0.476	1
Spearman	Average	0.818	0.781	0.790 (*)	0.783 (*)	0.769	5
	Ward's	0.864	0.780	0.767	0.759	0.773	5

	Complete	0.804	0.772	0.769	0.776	0.774(*)	5
	Single	0.639	0.624	0.671	0.630	0.632	0
Kendall's Tau	Average	0.101	0.048	0.011	-0.014	0.002	0
	Ward's	0.041	0.029	0.019	-0.013	0.001	0
	Complete	-0.270	-0.028	-0.006	-0.006	-0.001	0
	Single	0.101	0.032	-0.011	0.006	0.001	0
Winsorized	Average	0.830	0.798 (*)	0.770	0.709	0.723	5
	Ward's	0.891	0.715	0.708	0.712	0.708	5
	Complete	0.830	0.704	0.696	0.664	0.662	2
	Single	0.894(*)	0.797	0.660	0.602	0.566	2
Freq		11	10	7	7	7	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.2 shows the results of CCC for 16 clustering algorithms under normal distribution when $p = 3$. When $n < 100$, the largest coefficients were dominated by Winsorized, followed by Spearman, and Pearson. However, when $n > 100$, Spearman coefficients were more larger for average linkage, Ward's and complete linkage compared to Pearson and Winsorized. Kendall's Tau was not well performed for any sample size and linkage methods as it had the most lowest coefficients. When $n = 10$, *Pearson-Ward's* and *Winsorized-single* were the dominant algorithms. When $n = 50$, *Winsorized-average* had the largest coefficients. When $n = 100$ and 200, *Spearman-average* had the largest coefficients. When $n = 500$, *Spearman-complete* is the dominant algorithm. Overall, as for conclusion, when $p = 3$, the appropriate clustering algorithm is *Spearman-average* as it has the most frequency of having largest coefficients. Next, Table 4.3 displays the results of CCC for 16 clustering algorithms under normal distribution when $p = 5$.

Table 4.3

Average value of CCC for 16 clustering algorithms under normal distribution when number of variables, $p = 5$

Correlation -based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient (CCC)					
		$p = 5$					Freq
		Samples size. n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.744	0.647	0.598 (*)	0.538	0.504	1
	Ward's	0.752	0.624	0.534	0.518	0.475	1
	Complete	0.736	0.646	0.511	0.501	0.457	1
	Single	0.687	0.473	0.271	0.247	0.133	0
Spearman	Average	0.758	0.658(*)	0.584	0.552	0.552 (*)	1
	Ward's	0.767	0.619	0.595	0.539	0.521	1
	Complete	0.795 (*)	0.596	0.540	0.533	0.524	1
	Single	0.689	0.410	0.183	0.191	0.180	0
Kendall's Tau	Average	0.011	-0.009	0.001	0.004	0.002	0
	Ward's	-0.147	0.008	0.001	-0.005	-0.003	0
	Complete	0.025	0.004	0.007	0.013	-0.003	0
	Single	-0.010	0.011	-0.009	-0.002	-0.086	0
Winsorized	Average	0.750	0.644	0.563	0.543	0.521	1
	Ward's	0.753	0.653	0.562	0.553 (*)	0.472	1
	Complete	0.789	0.601	0.550	0.510	0.459	1
	Single	0.702	0.486	0.435	0.245	0.129	1
Freq		10	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

On the whole, Winsorized had the highest total frequency of CCC value more than 0.7, Spearman come second closed and the third was Pearson. This is shown by the numbers stated in the right-last column of the table. However, in terms of coefficient, Spearman remains the largest when $n = 10$ until 500 compared to Winsorized and Pearson. Kendall's Tau still did not perform well for every sample size and linkage methods as it has the lowest correlation values among all.

As for the linkage methods, average, Ward's and complete were well performed as they were mostly had high coefficient values. Meanwhile, single linkage is the worst as it mostly had the lowest coefficient values. However, CCC values for all of the correlation coefficients deteriorate severely as sample size increased. When $n = 10$,

the dominant algorithm are *Spearman-complete*. Additionally, when $n = 50$ and 500 , *Spearman-average* shows the largest coefficient. Then, when $n = 100$ and $n = 200$, the largest coefficients are *Pearson-average* and *Winsorized-Ward's*, respectively. In a nutshell, when $p = 5$, the appropriate clustering algorithm is *Spearman-average* as it had more frequency of having largest coefficient at sample size $n = 50$ and 500 .

Table 4.4

Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 10$

Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient (CCC)					
		$p = 10$					
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	Freq
Pearson	Average	0.791 (*)	0.506	0.456	0.409	0.375 (*)	1
	Ward's	0.625	0.491	0.447	0.380	0.315	0
	Complete	0.675	0.502	0.406	0.378	0.305	0
	Single	0.698	0.324	0.173	0.124	0.066	0
Spearman	Average	0.754	0.543 (*)	0.477 (*)	0.419	0.368	1
	Ward's	0.635	0.477	0.432	0.382	0.334	0
	Complete	0.669	0.481	0.436	0.396	0.331	0
	Single	0.680	0.374	0.273	0.162	0.068	0
Kendall's Tau	Average	-0.026	0.003	0.013	0.012	0.003	0
	Ward's	0.178	-0.046	0.010	0.003	0.010	0
	Complete	0.126	0.024	-0.007	0.009	0.012	0
	Single	0.116	0.026	-0.015	-0.006	0.009	0
Winsorized	Average	0.756	0.515	0.442	0.433 (*)	0.359	1
	Ward's	0.654	0.508	0.435	0.397	0.333	0
	Complete	0.688	0.501	0.417	0.378	0.318	0
	Single	0.655	0.290	0.196	0.140	0.059	0
Freq		3	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above

When $p = 10$, under normal distribution, Pearson, Spearman and Winsorized has the same one frequency of CCC above 0.7. Kendall's Tau did not perform well as usual in every sample size and linkage methods. As for the linkage method, average

outperformed others as it mostly had large coefficients. Complete come second and the third is Ward's method. Both complete linkage and Ward's method were having a very close coefficient. Nonetheless, when comparing Ward's and complete linkage, it shows that Ward's perform well when sample size is more than 100 as it mostly had slightly higher coefficients than complete. Single linkage remained the worst linkage with low coefficients. In addition, the coefficients decreased below 0.7 value for any condition as sample size increased except for the average linkage when $n = 10$. When $n = 10$, the dominant algorithm is *Pearson-average*. When $n = 50$ and 100, *Spearman-average* had the largest coefficients. When $n = 200$ and $n = 500$, *Winsorized-average* and *Pearson-average* had the largest coefficient. In summary, when $p = 10$, the appropriate clustering algorithm is *Pearson-average* as it had the most frequency of having largest coefficient at sample size, $n = 50$ and 100.

Table 4.5 shows the results of CCC for 16 clustering algorithms under normal distribution when $p = 20$. Overall, it can be seen that the CCC values for all conditions had start to fall below 0.7 under this number of variable. Starting from this section, the coefficient between 0.5 to 0.69 is used to determine the dominant distance.

Table 4.5

Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 20$

Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient (CCC)					
		$p = 20$					Freq
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.652	0.472	0.415	0.344	0.276	0
	Ward's	0.674	0.439	0.374	0.305	0.241	0
	Complete	0.688 (*)	0.404	0.362	0.287	0.228	0
	Single	0.556	0.302	0.130	0.105	0.055	0
Spearman	Average	0.672	0.488	0.435	0.376 (*)	0.278	0

Kendall's Tau	Ward's	0.655	0.493 (*)	0.385	0.296	0.246	0
	Complete	0.648	0.405	0.343	0.301	0.238	0
	Single	0.577	0.323	0.217	0.096	0.058	0
	Average	-0.037	-0.015	-0.011	-0.001	0.006	0
	Ward's	0.179	0.010	0.007	-0.012	0.002	0
	Complete	-0.060	-0.030	0.020	-0.018	0.009	0
	Single	-0.074	-0.038	-0.005	0.008	0.003	0
	Average	0.662	0.475	0.436 (*)	0.337	0.282	0
	Ward's	0.638	0.461	0.349	0.307	0.232	0
	Complete	0.670	0.437	0.322	0.285	0.232	0
Winsorized	Single	0.579	0.311	0.154	0.106	0.032	0
	Freq	0	0	0	0	0	

Note : : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Based on distances, Pearson, Spearman and Winsorized performed equally in terms of having the coefficients that was between 0.5 to 0.69. As for the linkage methods, average, Ward's and complete also performed almost equally. In addition, by comparing Ward's and complete linkage, it showed that Ward's perform well as it had higher coefficients than complete when sample size is more than 10 except for $n = 200$ under the Spearman distance (but the value was very close). Single linkage stayed as the worst linkage method in any situation. When $n = 10$, *Pearson-complete* linkage is the dominant algorithm. When $n = 50$, $n = 100$, and $n = 200$, the largest coefficients belongs to *Spearman-Ward's*, *Winsorized-average*, and *Spearman-average*, respectively. When $n = 500$, there is no such appropriate algorithm to be used as the coefficients are below 0.3. Overall, it can be concluded that the appropriate clustering algorithm is *Pearson-complete* as its coefficient was the highest among others.

Table 4.6

Average value of CCC for 16 clustering algorithms under normal distribution when the number of variables, $p = 30$

Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient (CCC)					
		$p = 30$					Freq
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.672	0.464	0.380	0.329	0.256	0
	Ward's	0.617	0.404	0.359	0.275	0.210	0
	Complete	0.611	0.389	0.357	0.274	0.175	0
	Single	0.682	0.206	0.146	0.084	0.034	0
Spearman	Average	0.647	0.469	0.372	0.340 (*)	0.262	0
	Ward's	0.674	0.415	0.340	0.266	0.209	0
	Complete	0.628	0.391	0.348	0.227	0.198	0
	Single	0.585	0.217	0.137	0.077	0.060	0
Kendall's Tau	Average	0.236	-0.031	-0.006	0.008	-0.002	0
	Ward's	0.374	0.021	-0.004	-0.002	-0.003	0
	Complete	-0.105	0.024	-0.005	-0.002	0.003	0
	Single	0.262	-0.022	0.025	0.009	-0.005	0
Winsorized	Average	0.638	0.473 (*)	0.390 (*)	0.319	0.254	0
	Ward's	0.698 (*)	0.429	0.332	0.265	0.203	0
	Complete	0.606	0.382	0.296	0.253	0.202	0
	Single	0.636	0.228	0.105	0.086	0.040	0
Freq		0	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.6 presents the average value of CCC for 16 clustering algorithms under normal distribution when $p = 30$. Overall, the highest frequency of CCC between 0.5 to 0.69 coefficients are dominated by Winsorized and Pearson, then followed by Spearman. Kendall's Tau remained have the worst correlation coefficient. In terms of linkage method, average, Ward's and complete are equally well performed. By comparing Ward's and complete, it displayed that Wards performed well as it frequently had slightly higher coefficients than complete. Single linkage did not perform well except for Pearson and Winsorized when $n = 10$ as their values were higher than complete linkage. When $n = 10, n = 50$ and 100, and n more than 100,

the dominant algorithms are *Winsorized-Ward's*, *Winsorized-average* and *Spearman-average* respectively. When $n = 500$, there is no such proper algorithm to be used. Overall, the appropriate algorithm is the pairing of *Winsorized-average* as it had more frequency of having the largest coefficients value.

Table 4.7

Summary of appropriate clustering algorithms under normal distribution

Number of variables, p	Distance-linkage	Range value of CCC
$p = 3$	<i>Spearman-average</i>	0.769 – 0.818
$p = 5$	<i>Spearman-average</i>	0.552 – 0.758
$p = 10$	<i>Pearson-average</i>	0.375 – 0.791
$p = 20$	<i>Pearson-complete</i>	0.362 – 0.688
$p = 30$	<i>Winsorized-average.</i>	0.319 – 0.638

As a summary of Table 4.2 until Table 4.6, Table 4.7 present the list of appropriate clustering algorithms according to the number of variables under normal distribution. It can be concluded that, when p increase, the value coefficient of cophenetic correlation is decreased. This table can be used as a guideline when deciding the suitable clustering algorithm. There is no single solution can be concluded as the suitability of the clustering algorithm depends on the number of variables and sample sizes.

4.2.2 Clustering Algorithm under Mild and Extreme Contaminated Distribution

In this section, the simulation results under mild and extreme contaminated distribution was presented. The analysis has been done for $p = 3, 5, 10, 20, 30$ and $n = 10, 50, 100, 200, 500$. In this simulation, the data was contaminated at two different level which are mild and extreme. For each case, the percentage of outliers

consist of 10%, 20%, and 30%. The statistical performance was present in the following tables.

Table 4.8

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu=3$) contaminated observations when the number of variables, $p = 3$

			Cophenetic correlation coefficient					
Conta- minated outlier (%)	Correlation -based distance measures	Clustering method (Linkage)	$p = 3$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n=200$	$n = 500$	Freq
10%	Pearson	Average	0.877(*)	0.762	0.752	0.749	0.727	5
		Ward's	0.811	0.718	0.732	0.716	0.700	5
		Complete	0.831	0.709	0.670	0.689	0.663	2
		Single	0.737	0.681	0.684	0.649	0.582	1
	Spearman	Average	0.873	0.777	0.778(*)	0.775	0.781(*)	5
		Ward's	0.768	0.739	0.771	0.771	0.750	5
		Complete	0.841	0.768	0.771	0.777(*)	0.773	5
		Single	0.623	0.628	0.633	0.635	0.631	0
	Kendall's Tau	Average	-0.079	-0.005	-0.017	-0.003	0.007	0
		Ward's	-0.046	-0.005	0.005	-0.005	-0.001	0
		Complete	-0.293	-0.010	-0.014	-0.005	0.002	0
		Single	-0.200	0.008	0.004	0.005	0.003	0
	Winsorized	Average	0.837	0.802(*)	0.758	0.748	0.718	5
		Ward's	0.845	0.787	0.717	0.719	0.711	5
		Complete	0.832	0.760	0.675	0.683	0.670	2
		Single	0.795	0.696	0.681	0.655	0.414	1
	Freq		11	9	7	7	7	
20%	Pearson	Average	0.885	0.795	0.734	0.726	0.714	5
		Wards	0.827	0.784	0.756	0.729	0.728	5
		Complete	0.817	0.755	0.680	0.684	0.676	2
		Single	0.781	0.681	0.641	0.665	0.617	1
	Spearman	Average	0.889(*)	0.793	0.778	0.779(*)	0.777(*)	5
		Wards	0.788	0.775	0.777	0.744	0.753	5
		Complete	0.874	0.762	0.780(*)	0.773	0.760	5
		Single	0.606	0.647	0.640	0.634	0.631	0
	Kendall's Tau	Average	0.101	0.048	0.011	-0.014	0.003	0
		Wards	-0.133	-0.006	0.015	0.008	0.001	0
		Complete	-0.037	0.039	0.002	0.002	-0.002	0
		Single	0.055	0.007	0.024	0.002	-0.001	0
	Winsorized	Average	0.885	0.797(*)	0.765	0.754	0.722	5
		Wards	0.800	0.721	0.743	0.742	0.708	5
		Complete	0.807	0.740	0.688	0.658	0.678	2

	Single	0.788	0.692	0.615	0.592	0.662	1	
Freq		11	9	7	7	7		
30%	Pearson	Average	0.895(*)	0.780(*)	0.768	0.733	0.721	5
		Wards	0.828	0.735	0.764	0.751	0.749	5
		Complete	0.809	0.710	0.659	0.656	0.663	2
		Single	0.792	0.606	0.646	0.653	0.584	1
	Spearman	Average	0.831	0.771	0.790(*)	0.784(*)	0.778(*)	5
		Wards	0.824	0.758	0.769	0.775	0.748	5
		Complete	0.812	0.750	0.771	0.780	0.773	5
		Single	0.606	0.625	0.628	0.632	0.631	0
	Kendall's Tau	Average	-0.228	-0.015	0.014	-0.002	0.004	0
		Wards	0.041	0.029	0.019	-0.013	0.005	0
		Complete	-0.030	0.014	0.011	-0.006	-0.002	0
		Single	0.099	-0.014	0.008	-0.002	-0.001	0
	Winsorized	Average	0.842	0.765	0.761	0.730	0.744	5
		Wards	0.834	0.711	0.732	0.736	0.721	5
		Complete	0.806	0.702	0.677	0.650	0.667	2
		Single	0.731	0.685	0.643	0.633	0.618	1
Freq		11	9	7	7	7		

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.8 shows the results of CCC for 16 clustering algorithms under normal distributions with mild contaminated observations with 10%, 20% and 30% of outliers injected when $p = 3$. As for the 10% contaminated observations, overall, the highest frequency of CCC above 0.7 are dominated by Spearman, followed by Winsorized and Pearson. This is shown by the amount stated at the right-last column of the table. By comparing Winsorized and Pearson, it showed that Winsorized had slightly more higher coefficients than Pearson even though their frequency were the same. Kendall's Tau did not worked well at any conditions. As for the linkage methods, average and Ward's performed well. Complete worked well when combined with Spearman as distance measure as it displayed high coefficients at all percentage of outliers. In addition, for Pearson, Spearman, and Winsorized, all the coefficients value decreased when sample size increased. It was portrayed by the number of bold values that changed from 11 to 9, and 7. When $n = 10$, $n = 50$, $n =$

100, $n = 200$ and $n = 500$, the dominant algorithms are *Pearson-average*, *Winsorized-average*, *Spearman-average*, *Spearman-complete*, and *Spearman-average*, respectively. As a conclusion, the appropriate clustering algorithm when there was 10% of outliers under mild contaminated observations is the combination of *Spearman-average* as it had the most frequency of having largest coefficients at $n = 100$ and 500.

As for the 20% contaminated observations, on the whole, the highest frequency of CCC that was more than 0.7 are dominated by Spearman, followed closely by Winsorized and Pearson. Kendall's Tau had the lowest coefficients at any conditions. As for the linkage methods, average and Ward's functioned very well. Complete also performed well specifically when combined with Spearman. When $n = 10$, $n = 50$, and $n = 100$, the dominant algorithm are *Spearman-average*, *Winsorized-average* and *Spearman-complete*, respectively. When $n = 200$ and 500, the prevalent algorithm is *Spearman-average*. Therefore, it can be concluded that the appropriate clustering algorithm is *Spearman-average* as it had the highest frequency of having largest coefficients at $n = 10, 200$ and 500.

At 30% contaminated observations, Spearman had the highest frequency of CCC above 0.7, Then, followed by Pearson and Winsorized. Kendall's Tau remained the worst distance measure. In terms of linkage methods, average and Ward's were equally well performed as they have close coefficients. Complete linkage functioned well when combined with Spearman. When $n = 10$ and 50, *Pearson-average* was the dominant algorithm. Meanwhile, when $n = 100$ and above, the prevalent algorithm was *Spearman-average*. As for the conclusion, the appropriate clustering algorithm when there is 30% of mild contaminated observations was also *Spearman-*

average as it's frequency of having largest coefficients was the highest at $n = 100, 200$ and 500 .

Table 4.9

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu=3$) contaminated observations when the number of variables, $p = 5$

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 5$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.754	0.639	0.572	0.548	0.509	1
		Ward's	0.737	0.582	0.550	0.509	0.491	1
		Complete	0.726	0.538	0.545	0.491	0.462	1
		Single	0.659	0.437	0.344	0.229	0.153	0
	Spearman	Average	0.762	0.625	0.603	0.565	0.565(*)	1
		Ward's	0.745	0.580	0.582	0.568(*)	0.555	1
		Complete	0.772	0.577	0.569	0.555	0.514	1
		Single	0.625	0.473	0.098	0.145	0.184	0
	Kendall's Tau	Average	0.105	0.035	-0.012	0.007	0.002	0
		Ward's	-0.071	-0.047	0.010	0.002	0.001	0
		Complete	-0.126	-0.022	0.032	-0.016	0.002	0
		Single	0.204	0.026	-0.008	0.008	0.003	0
	Winsorized	Average	0.734	0.667(*)	0.627(*)	0.550	0.524	1
		Ward's	0.773(*)	0.590	0.566	0.521	0.469	1
		Complete	0.761	0.606	0.560	0.487	0.452	1
		Single	0.676	0.407	0.412	0.353	0.162	0
	Freq		9	0	0	0	0	
20%	Pearson	Average	0.762	0.635	0.598	0.528	0.515	1
		Ward's	0.789	0.579	0.519	0.530	0.514	1
		Complete	0.778	0.595	0.558	0.522	0.457	1
		Single	0.617	0.393	0.449	0.278	0.165	0
	Spearman	Average	0.800(*)	0.647(*)	0.594	0.551(*)	0.580(*)	1
		Ward's	0.747	0.599	0.621(*)	0.543	0.525	1
		Complete	0.720	0.557	0.532	0.523	0.536	1
		Single	0.708	0.528	0.190	0.185	0.180	1
	Kendall's Tau	Average	-0.221	0.012	0.041	-0.015	-0.002	0
		Ward's	0.022	-0.008	-0.042	-0.006	-0.003	0
		Complete	-0.289	0.039	0.002	0.005	0.007	0
		Single	0.073	0.015	0.035	0.014	0.003	0
	Average		0.737	0.618	0.581	0.543	0.528	1

30%	Winsorized	Ward's	0.750	0.618	0.520	0.515	0.481	1
		Complete	0.721	0.577	0.546	0.511	0.457	1
		Single	0.639	0.582	0.312	0.211	0.134	0
		Freq	10	0	0	0	0	
	Pearson	Average	0.757	0.604	0.591	0.544	0.516	1
		Ward's	0.746	0.639	0.555	0.510	0.509	1
		Complete	0.722	0.541	0.539	0.512	0.449	1
		Single	0.685	0.484	0.314	0.173	0.179	0
	Spearman	Average	0.799(*)	0.632	0.591	0.564(*)	0.574(*)	1
		Ward's	0.784	0.639	0.593	0.532	0.536	1
		Complete	0.750	0.557	0.560	0.552	0.520	1
		Single	0.715	0.546	0.309	0.177	0.184	1
	Kendall's Tau	Average	0.045	-0.036	-0.011	0.002	-0.003	0
		Ward's	-0.195	-0.047	-0.003	-0.005	-0.006	0
		Complete	0.108	-0.060	-0.009	-0.015	0.004	0
		Single	0.087	0.012	0.007	-0.004	0.002	0
	Winsorized	Average	0.760	0.674(*)	0.604(*)	0.543	0.524	1
		Ward's	0.703	0.608	0.550	0.529	0.508	1
		Complete	0.751	0.543	0.559	0.486	0.452	1
		Single	0.637	0.596	0.334	0.357	0.199	0
	Freq		10	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.9 shows the results of CCC for 16 clustering algorithms under normal distributions, mild contaminated observations with 10%, 20% and 30% of outliers injected when $p = 5$. As for the 10% contaminated observations, overall, Spearman, Winsorized and Pearson had the same total frequency of CCC above 0.7. Kendall's Tau did not worked well at any conditions. In terms of linkage methods, average was the best linkage as it frequently had the highest coefficients. Ward's and complete performed almost equally in terms of having close CCC value. Nonetheless, by comparing Ward's and complete, it showed that Ward's functioned well when sample size was more than 50 as it often displayed higher CCC value than complete. Single linkage was the weakest linkage as it mostly had the lowest CCC value. Besides, for Pearson, Spearman, and Winsorized, all the coefficients decreased when sample size increased as shown by the number of bold values that changed from 9 to

0. Next, when $n = 10$, the largest coefficient is *Winsorized-Ward's*. When $n = 50$ and 100, *Winsorized-average* is the leading algorithm. When $n = 200$ and $n = 500$, the dominant algorithm are *Spearman-Ward's* and *Spearman-average*. In conclusion, the appropriate clustering algorithm when there was 10% of outliers when $p = 5$ is *Winsorized-average* as it had the highest frequency of having largest coefficients at $n = 50$ and 100.

As for the 20% of outliers injected for contaminated observations, overall, the total highest frequency of CCC that was more than 0.7 were dominated by Spearman. Winsorized and Pearson had the same total frequency of CCC above 0.7. Kendall's Tau had the lowest coefficients at any conditions. As for the linkage methods, average, Ward's and complete were equally well-performed as they were having high coefficients and their value was close. Additionally, it can be observed that *Spearman-average* was the most dominant algorithm at all sample sizes except when $n = 100$. When $n = 100$, *Spearman-Ward's* had the largest coefficient. Hence, the clustering algorithm that demonstrates superior performance when there is 20% of outliers is *Spearman-average*.

At 30% of contaminated observations, Spearman remained the first that had highest frequency of CCC above 0.7, then followed by Winsorized and Pearson. Kendall's Tau remained the worst distance measure. In terms of linkage methods, average and Ward's performed well equally in terms of having high coefficients (and their value was close), followed by complete. When $n = 10, 200$ and 500, *Spearman-average* was the prevalent algorithm. Meanwhile, $n = 50$ and 100, the dominant algorithm was *Winsorized-average*. As for the conclusion, the appropriate clustering algorithm

in the presence of 30% injected outliers for mildly contaminated observations was the *Spearman-average*.

Table 4.10

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu=3$) contaminated observations when the number of variables, $p = 10$.

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 10$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.728	0.561	0.475(*)	0.435(*)	0.364	1
		Ward's	0.687	0.475	0.448	0.403	0.330	0
		Complete	0.655	0.442	0.460	0.367	0.323	0
		Single	0.682	0.316	0.185	0.151	0.068	0
	Spearman	Average	0.729(*)	0.589(*)	0.469	0.423	0.385(*)	1
		Ward's	0.723	0.497	0.435	0.368	0.329	1
		Complete	0.650	0.447	0.474	0.363	0.335	0
		Single	0.651	0.396	0.260	0.089	0.089	0
	Kendall's Tau	Average	-0.297	0.007	0.021	-0.001	0.004	0
		Ward's	0.053	-0.050	-0.001	0.008	-0.001	0
		Complete	-0.071	-0.033	0.035	-0.006	-0.002	0
		Single	-0.032	0.027	0.002	-0.003	0.005	0
	Winsorized	Average	0.701	0.541	0.454	0.406	0.360	1
		Ward's	0.690	0.478	0.453	0.376	0.309	0
		Complete	0.685	0.477	0.421	0.367	0.314	0
		Single	0.636	0.378	0.230	0.108	0.099	0
	Freq		4	0	0	0	0	
20%	Pearson	Average	0.771	0.544	0.480 (*)	0.415 (*)	0.363	1
		Ward's	0.669	0.485	0.433	0.379	0.320	0
		Complete	0.679	0.483	0.430	0.390	0.312	0
		Single	0.607	0.380	0.229	0.125	0.066	0
	Spearman	Average	0.786(*)	0.544	0.478	0.405	0.381 (*)	1
		Ward's	0.676	0.501	0.438	0.397	0.342	0
		Complete	0.694	0.454	0.415	0.377	0.324	0
		Single	0.603	0.304	0.246	0.104	0.054	0
	Kendall's Tau	Average	-0.078	-0.005	-0.006	0.001	0.001	0
		Ward's	0.022	0.033	0.010	-0.002	0.003	0
		Complete	-0.330	-0.009	-0.005	-0.004	0.003	0
		Single	-0.239	0.055	-0.015	0.009	0.001	0
	Winsorized	Average	0.762	0.580(*)	0.469	0.411	0.370	1
		Ward's	0.739	0.469	0.425	0.392	0.321	1

	Complete	0.657	0.497	0.410	0.358	0.309	0	
	Single	0.578	0.277	0.137	0.140	0.066	0	
	Freq	4	0	0	0	0		
30%	Pearson	Average	0.700	0.538	0.473	0.412	0.363(*)	1
		Ward's	0.705	0.496	0.429	0.372	0.324	1
		Complete	0.661	0.477	0.424	0.388	0.320	0
		Single	0.640	0.334	0.225	0.186	0.067	0
	Spearman	Average	0.778(*)	0.553(*)	0.456	0.432(*)	0.357	1
		Ward's	0.655	0.488	0.458	0.406	0.322	0
		Complete	0.658	0.474	0.433	0.369	0.324	0
		Single	0.649	0.267	0.234	0.112	0.069	0
	Kendall's Tau	Average	0.258	0.007	0.033	-0.003	0.003	0
		Ward's	0.156	0.001	-0.030	-0.012	0.004	0
		Complete	0.087	-0.030	-0.009	-0.001	-0.007	0
		Single	0.132	0.036	-0.025	0.003	0.003	0
	Winsorized	Average	0.719	0.543	0.492(*)	0.407	0.360	1
		Ward's	0.626	0.510	0.451	0.394	0.320	0
		Complete	0.642	0.482	0.418	0.393	0.318	0
		Single	0.609	0.394	0.228	0.142	0.079	0
	Freq	4	0	0	0	0		

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

The results of CCC for 16 clustering algorithms under normal distributions, mild contamination with 10%, 20%, and 30% injected outliers when $p = 10$, are displayed in Table 4.10. As for the 10% contaminated observations, Spearman had the highest frequency of CCC values above 0.7, followed by Pearson and Winsorized. Kendall's Tau still did not perform satisfactorily at any conditions. In terms of linkage methods, average outperformed others as it frequently displayed higher coefficients followed closely by Ward's. Complete linkage third. In addition, for Pearson, Spearman, and Winsorized, all the coefficients decreased severely when sample size increased. When $n = 10$, 50 and 500, *Spearman-average* was the leading algorithm while when , $n = 100$ and 200, *Pearson-average* had the largest coefficients. Thus, the appropriate clustering algorithm for $p = 10$, with a presence of 10% outliers, was *Spearman-average*.

Next, as for the 20% contaminated observations, on the whole, the highest frequency of CCC that is more than 0.7 is dominated by Winsorized. Spearman and Pearson had only one CCC value that is above 0.7. Kendall's Tau remained the weakest coefficient. As for the linkage methods, average remained the best while Ward's and complete are equivalently well-performed as they were having close coefficients. When $n = 10$ and 500, the dominant algorithm was *Spearman-average*, while when $n = 50$ the largest coefficient was *Winsorized-average*. When $n = 100$ and 200, the leading algorithm was *Pearson-average*. Therefore, it can be concluded that the appropriate clustering algorithm when there was a presence of 20% outliers when $p = 10$ was *Spearman-average* as it had the one of the highest coefficients especially at $n = 10$ even though it had the same frequency of having largest coefficients as *Pearson-average*.

As for the presence of 30% outliers, Pearson had the highest frequency of CCC above 0.7. Spearman and Winsorized had only one frequency of CCC above 0.7 while Kendall's Tau had none. In terms of linkage methods, average was the best followed by Ward's and complete. When $n = 10$, 50 and 200, *Spearman-average* had the largest coefficients. Meanwhile, when $n = 100$ and $n = 500$, the prevalent algorithms were *Winsorized-average* and *Pearson-average*. Hence, the appropriate clustering algorithm when there was 30% of outliers was *Spearman-average*.

Table 4.11

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu=3$) contaminated observations when the number of variables, $p = 20$.

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 20$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.694(*)	0.494(*)	0.405	0.353	0.283	0
		Ward's	0.642	0.444	0.384	0.295	0.243	0
		Complete	0.677	0.422	0.336	0.296	0.212	0
		Single	0.595	0.308	0.169	0.100	0.053	0
	Spearman	Average	0.649	0.486	0.416(*)	0.357(*)	0.288(*)	0
		Ward's	0.630	0.464	0.383	0.316	0.238	0
		Complete	0.628	0.402	0.375	0.297	0.235	0
		Single	0.612	0.274	0.193	0.099	0.060	0
	Kendall's Tau	Average	-0.029	-0.036	-0.008	-0.005	0.003	0
		Ward's	-0.008	0.013	-0.016	-0.004	-0.001	0
		Complete	-0.060	-0.009	-0.005	0.016	-0.002	0
		Single	-0.194	0.013	0.019	0.005	-0.005	0
	Winsorized	Average	0.674	0.452	0.403	0.349	0.283	0
		Ward's	0.651	0.470	0.394	0.306	0.226	0
		Complete	0.605	0.409	0.366	0.301	0.229	0
		Single	0.508	0.209	0.231	0.103	0.045	0
	Freq		0	0	0	0	0	
20%	Pearson	Average	0.671	0.487(*)	0.410	0.352	0.283	0
		Ward's	0.645	0.433	0.351	0.317	0.234	0
		Complete	0.659	0.410	0.345	0.283	0.234	0
		Single	0.483	0.325	0.165	0.087	0.069	0
	Spearman	Average	0.665	0.470	0.421(*)	0.353	0.287	0
		Ward's	0.611	0.409	0.391	0.333	0.253	0
		Complete	0.625	0.399	0.332	0.293	0.251	0
		Single	0.563	0.208	0.150	0.095	0.049	0
	Kendall's Tau	Average	0.053	-0.018	-0.012	-0.002	0.002	0
		Ward's	0.077	-0.020	-0.015	-0.003	-0.001	0
		Complete	-0.215	0.027	-0.005	0.013	-0.004	0
		Single	0.148	-0.022	-0.015	0.007	-0.005	0
	Winsorized	Average	0.672(*)	0.486	0.402	0.354(*)	0.290(*)	0
		Ward's	0.641	0.466	0.377	0.306	0.238	0
		Complete	0.641	0.473	0.367	0.274	0.208	0
		Single	0.523	0.358	0.165	0.106	0.051	0
	Freq		0	0	0	0	0	
30%	Pearson	Average	0.699(*)	0.452	0.404	0.342	0.286	0

	Ward's	0.654	0.441	0.366	0.298	0.241	0
	Complete	0.698	0.453	0.345	0.293	0.232	0
	Single	0.552	0.237	0.149	0.078	0.039	0
Spearman	Average	0.664	0.488(*)	0.392	0.355(*)	0.292(*)	0
	Ward's	0.622	0.462	0.377	0.315	0.242	0
	Complete	0.642	0.416	0.358	0.291	0.214	0
	Single	0.557	0.245	0.215	0.128	0.063	0
Kendall's Tau	Average	0.111	-0.026	-0.008	-0.002	-0.004	0
	Ward's	-0.059	0.055	0.016	0.005	-0.006	0
	Complete	-0.115	-0.006	0.006	0.011	0.001	0
	Single	-0.068	0.046	-0.016	0.004	0.001	0
Winsorized	Average	0.662	0.475	0.417(*)	0.340	0.281	0
	Ward's	0.633	0.394	0.367	0.293	0.244	0
	Complete	0.602	0.434	0.332	0.272	0.226	0
	Single	0.617	0.227	0.091	0.112	0.053	0
Freq		0	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

The results of CCC for 16 clustering algorithms under normal distributions with injected outliers of 10%, 20%, and 30% when $p = 20$ are illustrated in Table 4.11. Overall, under this number of variable, all the CCC values was below 0.7 for any conditions. Starting from this section, the coefficient between 0.5 to 0.69 was used to determine the dominant distance.

As for 10% of injected outliers, Spearman had the highest frequency of having higher CCC between 0.5 to 0.69 coefficients especially at $n = 10$, then followed by Pearson and Winsorized. Kendall's Tau was the worst at any conditions. In terms of linkage methods, average performed better than others as it often had the highest coefficients. Then, followed by Ward's and complete. Single linkage was the weakest. The dominant algorithms when $n = 10$ and 50 was *Pearson-average* while when $n = 100$ and 200 was *Spearman-average*. When $n = 500$, there was no such appropriate algorithm to be used as the coefficients were below 0.3. As a conclusion, the appropriate clustering algorithms when there was a presence of 10% outliers when $p = 20$ was *Pearson-average* as it had higher coefficients than

Spearman-average although they were having equal frequency of having largest coefficients.

As for the 20% contaminated observations, overall, based on distance, Spearman and Winsorized performed equally in terms of having high CCC between 0.5 to 0.69, especially at $n = 10$. Pearson comes third. As for the linkage methods, average performed the best as it often had the highest coefficients followed closely by Ward's. Complete was the third while single linkage was the worst. When $n = 10$ and 200, $n = 50$, and $n = 100$, the dominant algorithm are *Winsorized-average*, *Pearson -average* and *Spearman-average*, respectively. When $n = 500$, there was no proper algorithm. Therefore, it can be concluded that the appropriate clustering algorithm was the combination of *Winsorized- average*.

At 30% contaminated observations, it can be observed that Winsorized had the highest frequency of CCC above 0.6. Pearson and Spearman had the same frequency of CCC between 0.5 to 0.69. Kendall's Tau has none. In terms of linkage methods, average was the best while, Ward's and complete were equivalently functioned well. By comparing Ward's and complete, complete performed well when sample size, n is 10 under Pearson and Spearman and $n = 50$ under Pearson and Winsorized distances as they displayed higher coefficients than Ward's. When $n = 10$, $n = 50$ and 200, and $n = 100$, the prevalent algorithms were *Pearson-average*, *Spearman-average* and *Winsorized-average*, respectively. As for the conclusion, the appropriate clustering algorithm when there was 30% of mild contaminated observations when $p = 20$ was *Spearman- average*.

Table 4.12

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for mild ($\mu=3$) contaminated observations when the number of variables, $p = 30$.

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 30$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.646	0.479	0.406(*)	0.330	0.255	0
		Ward's	0.614	0.411	0.358	0.292	0.213	0
		Complete	0.660	0.389	0.314	0.276	0.189	0
		Single	0.537	0.255	0.163	0.094	0.040	0
	Spearman	Average	0.670	0.447	0.384	0.340(*)	0.257	0
		Ward's	0.640	0.453	0.346	0.284	0.205	0
		Complete	0.662	0.419	0.306	0.233	0.183	0
		Single	0.618	0.220	0.170	0.103	0.038	0
	Kendall's Tau	Average	0.257	-0.020	0.004	0.007	0.003	0
		Ward's	0.191	-0.064	0.003	-0.003	-0.004	0
		Complete	0.229	0.021	-0.006	0.004	0.001	0
		Single	0.206	0.006	0.017	-0.015	0.002	0
	Winsorized	Average	0.674(*)	0.480(*)	0.401	0.332	0.266	0
		Ward's	0.617	0.402	0.354	0.279	0.207	0
		Complete	0.635	0.391	0.356	0.243	0.197	0
		Single	0.593	0.273	0.179	0.125	0.041	0
	Freq		0	0	0	0	0	
20%	Pearson	Average	0.636	0.485(*)	0.377	0.329(*)	0.252	0
		Ward's	0.644	0.448	0.346	0.269	0.207	0
		Complete	0.613	0.413	0.339	0.241	0.198	0
		Single	0.540	0.216	0.185	0.100	0.039	0
	Spearman	Average	0.665	0.456	0.385	0.328	0.254	0
		Ward's	0.655	0.427	0.359	0.263	0.209	0
		Complete	0.620	0.376	0.306	0.259	0.192	0
		Single	0.608	0.240	0.164	0.098	0.046	0
	Kendall's Tau	Average	-0.305	-0.022	-0.026	0.008	0.002	0
		Ward's	0.269	0.029	0.006	0.004	-0.004	0
		Complete	0.129	-0.043	-0.009	-0.002	0.005	0
		Single	0.180	-0.002	0.017	0.003	0.003	0
	Winsorized	Average	0.655	0.465	0.405(*)	0.314	0.259	0
		Ward's	0.669(*)	0.412	0.335	0.277	0.211	0
		Complete	0.603	0.386	0.302	0.279	0.178	0
		Single	0.560	0.218	0.152	0.079	0.044	0
	Freq		0	0	0	0	0	
Average		0.652	0.438	0.388	0.330	0.257	0	

30%	Pearson	Ward's	0.623	0.420	0.336	0.286	0.203	0
		Complete	0.613	0.369	0.305	0.265	0.193	0
		Single	0.549	0.265	0.148	0.094	0.050	0
	Spearman	Average	0.669(*)	0.461(*)	0.393	0.330	0.264	0
		Ward's	0.608	0.410	0.314	0.273	0.210	0
		Complete	0.633	0.339	0.335	0.249	0.175	0
		Single	0.600	0.239	0.197	0.098	0.048	0
	Kendall's Tau	Average	0.270	0.035	0.005	0.005	-0.002	0
		Ward's	-0.049	-0.033	0.007	0.003	-0.002	0
		Complete	-0.143	0.012	0.011	-0.006	-0.005	0
		Single	-0.006	-0.023	-0.017	0.006	-0.003	0
	Winsorized	Average	0.644	0.457	0.398(*)	0.346(*)	0.259	0
		Ward's	0.644	0.423	0.343	0.290	0.202	0
		Complete	0.612	0.371	0.306	0.271	0.193	0
		Single	0.539	0.268	0.159	0.077	0.045	0
	Freq		0	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

When $p = 30$, as for the 10% contaminated observations, overall, the highest frequency of CCC between 0.5 to 0.69 were dominated by Spearman, specifically at $n = 10$, then followed by Winsorized and Pearson. Kendall's Tau did not worked well as anticipated. In terms of linkage methods, average was the best linkage as it frequently had the highest coefficients than others. Ward's and complete performed equally in terms of having moderate coefficients between 0.3 to 0.699 while single linkage was the weakest. By comparing Ward's and complete, it showed that Ward's function well when sample size was more than 10. Next, when $n = 10$ and 50 the largest coefficient is *Winsorized-average*. When $n = 100$ and $n = 200$, the largest coefficients belongs to *Pearson-average* and *Spearman-average*, separately. When $n = 500$, there was no acceptable algorithm. In conclusion , the appropriate clustering algorithm when there was 10% of outliers at $p = 30$ was *Winsorized-average* as it had the highest frequency of having largest coefficients, particularly at $n = 10$ and 50.

As for the 20% of outliers injected for contaminated observations, overall, Spearman had the highest frequency of CCC for above 0.6. Then, followed by Winsorized and Pearson as they had the same frequency of CCC between 0.5 to 0.69. Kendall's Tau had none. As for the linkage methods, average was the best, followed by Ward's and third was complete. Single linkage came last. When $n = 10, n = 50, n = 100$ and $n = 200$, the dominant algorithms were, *Winsorized-Wards*, *Pearson-average*, *Winsorized-average* and *Pearson-average*, respectively. Hence, *Pearson-average* was the appropriate clustering algorithm.

At 30% of contaminated observations, based on distances, Spearman remained the first. Pearson and Winsorized have the same frequency of CCC between 0.5 to 0.69. Kendall's Tau remained the worst distance measure. In terms of linkage methods, average was the best as it always exhibited the highest coefficients, followed closely by and Ward's and complete. When $n = 10$ and 50, *Spearman-average* was the prevalent algorithm. Meanwhile, when $n = 100$ and 200, *Winsorized-average* takes the lead. As for conclusion, the appropriate clustering algorithm in the presence of 30% injected outliers for mildly contaminated observations when $p = 30$ was *Spearman-average* as it had higher coefficients than *Winsorized-average* although they had the same frequency of having largest coefficients.

Table 4.13

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu=5$) contaminated observations when the number of variables, $p = 3$.

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 3$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.854	0.771	0.741	0.723	0.717	5
		Ward's	0.849	0.744	0.755	0.720	0.691	4

20%		Complete	0.813	0.713	0.704	0.671	0.671	3
		Single	0.791	0.642	0.618	0.621	0.588	1
	Spearman	Average	0.857(*)	0.786	0.780(*)	0.777(*)	0.776(*)	5
		Ward's	0.850	0.754	0.760	0.772	0.754	5
		Complete	0.843	0.765	0.768	0.776	0.773	5
		Single	0.667	0.624	0.627	0.631	0.632	0
	Kendall's Tau	Average	0.048	0.025	0.002	0.007	-0.002	0
		Ward's	0.007	-0.005	-0.005	0.003	0.002	0
		Complete	-0.050	-0.016	0.003	-0.007	-0.003	0
		Single	0.109	0.043	-0.009	-0.010	-0.003	0
	Winsorized	Average	0.856	0.805(*)	0.754	0.730	0.721	5
		Ward's	0.823	0.792	0.776	0.718	0.705	5
		Complete	0.798	0.742	0.695	0.666	0.661	2
		Single	0.777	0.706	0.636	0.644	0.567	2
	Freq		11	10	8	7	6	
	Pearson	Average	0.855(*)	0.763	0.741	0.747	0.734	5
		Ward's	0.837	0.743	0.752	0.743	0.703	5
		Complete	0.826	0.721	0.696	0.678	0.679	2
		Single	0.807	0.640	0.642	0.637	0.549	1
30%	Spearman	Average	0.845	0.805	0.793(*)	0.778(*)	0.778(*)	5
		Ward's	0.830	0.807(*)	0.770	0.772	0.773	5
		Complete	0.829	0.798	0.772	0.772	0.775	5
		Single	0.648	0.620	0.637	0.633	0.633	0
	Kendall's Tau	Average	0.101	0.048	0.011	0.004	0.003	0
		Ward's	-0.069	0.008	0.002	-0.005	0.002	0
		Complete	-0.051	-0.034	0.009	-0.003	0.002	0
		Single	0.178	-0.018	0.001	0.005	0.004	0
	Winsorized	Average	0.854	0.752	0.750	0.708	0.707	5
		Ward's	0.853	0.714	0.756	0.714	0.707	5
		Complete	0.822	0.701	0.653	0.668	0.673	2
		Single	0.807	0.692	0.689	0.639	0.665	1
	Freq		11	9	7	7	7	
	Pearson	Average	0.836	0.755	0.744	0.723	0.715	5
		Ward's	0.820	0.751	0.725	0.712	0.716	5
		Complete	0.816	0.724	0.698	0.677	0.665	2
		Single	0.802	0.685	0.600	0.638	0.483	1
	Spearman	Average	0.820	0.814(*)	0.789(*)	0.776(*)	0.779(*)	5
		Ward's	0.859(*)	0.807	0.743	0.773	0.774	5
		Complete	0.824	0.792	0.771	0.773	0.764	5
		Single	0.647	0.624	0.627	0.628	0.632	0
	Kendall's Tau	Average	-0.228	-0.015	0.014	-0.002	0.004	0
		Ward's	0.165	0.007	-0.027	-0.001	0.007	0
		Complete	0.042	-0.070	-0.004	-0.004	0.002	0
		Single	-0.349	-0.007	-0.007	-0.006	-0.002	0

Winsorized	Average	0.848	0.763	0.752	0.732	0.725	5
	Ward's	0.821	0.764	0.773	0.758	0.720	5
	Complete	0.826	0.733	0.689	0.680	0.657	2
	Single	0.787	0.688	0.660	0.627	0.641	1
Freq		11	9	7	7	7	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

The results of CCC for 16 clustering algorithms under normal distributions, extreme contamination with 10%, 20%, and 30% injected outliers when $p = 3$, are displayed in Table 4.13. At 10% contaminated observations, overall, the highest frequency of CCC above 0.7 were dominated by Spearman, followed by Winsorized and third Pearson. This was shown by the numbers stated at the right-last column of the table. Kendall's Tau did not worked well at any conditions. In terms of linkage methods, average was the best linkage as it often displayed highest coefficients followed closely by Ward's and then complete. By comparing Ward's and complete, it showed that complete functioned well when sample size was more than 10 when combined with Spearman distance. Single linkage was the weakest linkage. Besides, for Pearson, Spearman, and Winsorized, all the coefficients decreased when sample size increased as shown by the number of bold values that changed from 11, 10, 8,7 and 6. Next, when $n = 10, 100, 200$ and 500, the largest coefficient was dominated by *Spearman-average*. When $n = 50$, *Winsorized-average* had the largest coefficient. In conclusion , the appropriate clustering algorithm when there was 10% of outliers when $p = 3$ under extreme contamination was *Spearman-average*.

As for the 20% of outliers injected, overall, the highest frequency of CCC that was more than 0.7 were dominated by Spearman, followed closely by Winsorized and Pearson. Kendall's Tau had the lowest coefficients at any conditions. As for the linkage methods, average and Ward's equally well-performed as they were displaying high coefficients and their values were close. Complete came third. Single

linkage had the lowest coefficients in any cases. When $n = 10$ and $n = 50$, *Pearson-average* and *Spearman-Ward's* were having the largest coefficients. When $n = 100$ and above, *Spearman-average* was the dominant algorithm. Hence, the clustering algorithm that demonstrates superior performance was *Spearman-average*.

At 30% of contaminated observations, Spearman remained the first that has highest frequency of CCC above 0.7, followed by Pearson and Winsorized. Kendall's Tau remained the worst distance measure. In terms of linkage methods, average and Ward's performed well equally, then followed by complete. When $n = 10$, *Spearman-Ward's* was the dominant algorithm. Meanwhile, $n = 50$ and above, the dominant algorithm was *Spearman-average*. As for the conclusion, the clustering algorithm that was appropriate in the presence of 30% injected outliers was *Spearman-average*.

Table 4.14

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu=5$) contaminated observations when the number of variables, $p = 5$.

Contami -nated outlier (%)	Correlation -based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					
			$p = 5$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	Freq
10%	Pearson	Average	0.774	0.630	0.571	0.543	0.503	1
		Ward's	0.750	0.604	0.534	0.543	0.495	1
		Complete	0.747	0.569	0.519	0.484	0.460	1
		Single	0.656	0.545	0.420	0.245	0.185	0
	Spearman	Average	0.788(*)	0.658(*)	0.620(*)	0.563(*)	0.553(*)	1
		Ward's	0.748	0.603	0.593	0.532	0.536	1
		Complete	0.725	0.550	0.559	0.513	0.525	1
		Single	0.673	0.442	0.140	0.188	0.185	0
	Kendall's Tau	Average	-0.192	-0.021	0.001	-0.012	-0.004	0
		Ward's	-0.152	0.010	-0.006	-0.008	0.003	0
		Complete	-0.329	0.014	0.012	-0.001	0.002	0
		Single	0.175	-0.069	-0.005	0.005	0.005	0
	Average	Average	0.787	0.641	0.574	0.545	0.519	1

20%	Winsorized	Ward's	0.757	0.606	0.566	0.502	0.484	1
		Complete	0.711	0.593	0.533	0.486	0.455	1
		Single	0.692	0.492	0.369	0.212	0.176	0
	Freq		9	0	0	0	0	
	Pearson	Average	0.787	0.618	0.606(*)	0.543	0.519	1
		Ward's	0.741	0.617	0.581	0.503	0.486	1
		Complete	0.720	0.551	0.530	0.488	0.450	1
		Single	0.676	0.497	0.305	0.193	0.152	0
	Spearman	Average	0.788(*)	0.619(*)	0.603	0.579(*)	0.546	1
		Ward's	0.745	0.618	0.548	0.545	0.560(*)	1
		Complete	0.759	0.605	0.543	0.539	0.530	1
		Single	0.675	0.520	0.214	0.181	0.178	0
	Kendall's Tau	Average	0.471	0.046	-0.028	0.006	-0.004	0
		Ward's	-0.092	-0.011	0.024	-0.002	-0.006	0
		Complete	-0.016	-0.062	0.010	0.001	0.005	0
		Single	0.039	0.013	0.005	0.003	-0.003	0
30%	Winsorized	Average	0.780	0.618	0.585	0.549	0.513	1
		Ward's	0.784	0.602	0.556	0.511	0.482	1
		Complete	0.712	0.570	0.537	0.485	0.450	1
		Single	0.695	0.403	0.303	0.229	0.188	0
	Freq		9	0	0	0	0	
	Pearson	Average	0.760	0.613	0.569	0.543	0.516	1
		Ward's	0.757	0.588	0.525	0.533	0.472	1
		Complete	0.727	0.569	0.512	0.496	0.455	1
		Single	0.671	0.389	0.346	0.233	0.117	0
	Spearman	Average	0.812(*)	0.666(*)	0.597(*)	0.573(*)	0.560(*)	1
		Ward's	0.726	0.600	0.588	0.540	0.531	1
		Complete	0.712	0.597	0.566	0.556	0.518	1
		Single	0.650	0.376	0.178	0.181	0.186	0
	Kendall's Tau	Average	0.002	-0.065	0.004	0.010	-0.003	0
		Ward's	-0.137	-0.021	-0.014	0.003	0.007	0
		Complete	0.054	-0.036	-0.007	-0.002	-0.005	0
		Single	0.220	-0.003	-0.018	-0.003	0.003	0
	Winsorized	Average	0.770	0.658	0.596	0.527	0.507	1
		Ward's	0.763	0.632	0.516	0.547	0.499	1
		Complete	0.703	0.584	0.551	0.500	0.457	1
		Single	0.694	0.324	0.385	0.237	0.235	0
	Freq		9	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

The results of CCC for 16 clustering algorithms under normal distribution, extreme contamination with 10%, 20%, and 30% injected outliers when $p = 5$, are displayed

in Table 4.14. As for the 10% contaminated observations, Pearson, Spearman and Winsorized were equivalently well-performed. Kendall's Tau did not perform satisfactorily at any conditions. In terms of linkage methods, average outperformed others, followed by Ward's and complete. Single linkage did not work effectively in most conditions. In addition, for Pearson, Spearman, and Winsorized, all the coefficients decreased severely when sample size increased. This was shown by the bold values number that changed from 9 to 0. Overall, it can be seen that *Spearman-average* was the appropriate clustering algorithm when there was 10% of outliers as it had the largest coefficients for all the sample sizes.

Next, as for the 20% contaminated observations, Spearman, Pearson and Winsorized performed well equally. Kendall's Tau was the weakest. As for the linkage methods, average remained the best while Ward's and complete were equivalently well-performed as they were having close coefficients. When $n = 10, 50$, and 200 the dominant algorithm were *Spearman-average*, while when $n = 100$ and $n = 500$, *Pearson-average* and *Spearman-Ward's* had the largest coefficients, respectively. Therefore, it can be concluded that the clustering algorithm that exhibit superior performance was *Spearman-average*.

As for the presence of 30% outliers, Pearson, Spearman, and Winsorized were equally well-performed while Kendall's Tau did not function well. In terms of linkage methods, average remained the best. Ward's came second followed closely by complete. Single linkage stayed the worst. In addition, at all sample sizes, *Spearman-average* had the largest coefficients. Hence, the appropriate clustering algorithm when there was 30% of outliers for $p = 5$ was *Spearman-average*.

Table 4.15

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu=5$) contaminated observations when the number of variables, $p = 10$.

Contaminated outlier (%)	Correlation -based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 10$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.793(*)	0.563(*)	0.457	0.424(*)	0.364	1
		Ward's	0.646	0.488	0.445	0.377	0.318	0
		Complete	0.618	0.499	0.423	0.382	0.315	0
		Single	0.644	0.374	0.190	0.125	0.080	0
	Spearman	Average	0.747	0.523	0.489(*)	0.423	0.376(*)	1
		Ward's	0.688	0.510	0.410	0.385	0.328	0
		Complete	0.677	0.478	0.440	0.387	0.325	0
		Single	0.601	0.288	0.201	0.157	0.074	0
	Kendall's Tau	Average	-0.191	0.023	0.008	0.004	0.002	0
		Ward's	-0.108	-0.017	-0.010	0.006	-0.001	0
		Complete	-0.273	0.081	-0.021	0.013	-0.002	0
		Single	-0.092	-0.003	-0.023	0.007	0.005	0
	Winsorized	Average	0.701	0.550	0.463	0.421	0.358	1
		Ward's	0.660	0.492	0.426	0.374	0.321	0
		Complete	0.661	0.475	0.445	0.354	0.324	0
		Single	0.643	0.355	0.243	0.134	0.079	0
	Freq		3	0	0	0	0	
20%	Pearson	Average	0.724	0.539	0.460	0.414	0.363	1
		Ward's	0.551	0.472	0.414	0.400	0.331	0
		Complete	0.599	0.468	0.441	0.378	0.327	0
		Single	0.560	0.353	0.218	0.146	0.078	0
	Spearman	Average	0.718	0.576(*)	0.490(*)	0.419	0.372(*)	1
		Ward's	0.693	0.508	0.428	0.393	0.315	0
		Complete	0.692	0.476	0.426	0.391	0.334	0
		Single	0.635	0.437	0.176	0.122	0.107	0
	Kendall's Tau	Average	-0.005	0.008	0.006	-0.003	-0.001	0
		Ward's	0.024	0.005	0.010	-0.003	0.003	0
		Complete	0.114	0.019	0.020	0.005	0.004	0
		Single	-0.199	-0.040	-0.006	-0.005	0.002	0
	Winsorized	Average	0.747(*)	0.565	0.472	0.431(*)	0.359	1
		Ward's	0.664	0.510	0.425	0.363	0.323	0
		Complete	0.678	0.489	0.442	0.378	0.315	0
		Single	0.570	0.328	0.225	0.136	0.074	0
	Freq		3	0	0	0	0	
30%	Average	0.732	0.535	0.510(*)	0.417(*)	0.352	1	

Pearson	Ward's	0.660	0.484	0.428	0.389	0.322	0
	Complete	0.623	0.468	0.412	0.393	0.312	0
	Single	0.595	0.434	0.215	0.138	0.079	0
Spearman	Average	0.730	0.528	0.477	0.410	0.382(*)	1
	Ward's	0.671	0.473	0.469	0.374	0.332	0
	Complete	0.656	0.499	0.434	0.382	0.326	0
	Single	0.607	0.285	0.134	0.162	0.062	0
Kendall's Tau	Average	0.067	0.022	0.009	0.010	0.002	0
	Ward's	-0.071	0.010	-0.006	0.009	0.005	0
	Complete	-0.073	-0.041	0.002	0.005	-0.002	0
	Single	0.122	-0.026	-0.005	0.012	-0.001	0
Winsorized	Average	0.806(*)	0.536(*)	0.466	0.396	0.355	1
	Ward's	0.604	0.502	0.424	0.415	0.312	0
	Complete	0.643	0.466	0.412	0.372	0.327	0
	Single	0.626	0.300	0.254	0.133	0.084	0
Freq		3	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.15 shows the results of CCC for 16 clustering algorithms under normal distributions, extreme contaminated observations with 10%, 20% and 30% of outliers injected when $p = 10$. At 10% contaminated observations, overall, Pearson, Spearman and Winsorized had only one frequency of CCC above 0.7 at average linkage and $n = 10$ for each distance. Kendall's Tau had none. As for the linkage methods, average is the best in terms of having higher coefficient than others. Meanwhile, Ward's and complete performed equally as they were having close coefficients. When $n = 10, 50$ and 200 the dominant algorithm were *Pearson-average*. When $n = 100$ and 500 , *Spearman-average* was the dominant algorithm. As a conclusion, the appropriate clustering algorithm was *Pearson-average*.

At 20% contaminated observations, on the whole, Pearson, Spearman and Winsorized had the same one frequency of CCC that was more than 0.7. Kendall's Tau, in contrast had none at any conditions. As for the linkage methods, average functioned the best. Ward's and complete functioned at the same level but when

being compared, mostly, complete excelled than Ward's at $n = 10$ and 100 for Pearson and Winsorized distances. Single linkage did not functioned effectively. When $n = 10$, and 200 , *Winsorized-average* had the largest coefficients. Meanwhile, when $n = 50$, 100 and 500 , the dominant algorithm were *Spearman-average*. Therefore, it can be concluded that the appropriate clustering algorithm was *Spearman- average*.

At 30% contaminated observations, Pearson, Spearman and Winsorized also had only one frequency of CCC above 0.7 at average linkage and $n = 10$ for each distance. Kendall's Tau remained the worst distance measure. In terms of linkage methods, average remained the first well-performed linkage. Ward's and complete functioned well equivalently as they were having close coefficients. When $n = 10$ and 50 , *Winsorized-average* was the dominant algorithm. When $n = 100$ and 200 , the prevalent algorithm was *Pearson-average*. When $n = 500$, *Spearman-average* had the largest coefficient. As for the conclusion, the appropriate clustering algorithm when there was 30% of extreme contaminated observations is *Winsorized-average* as it had higher coefficients than *Pearson-Average* although they had the same frequency of having largest coefficients.

Table 4.16

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu=5$) contaminated observations when the number of variables, $p = 20$.

Conta- minated outlier (%)	Correlation -based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					
			$p = 20$					Freq
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.685	0.522(*)	0.408	0.345	0.281	0
		Ward's	0.696	0.442	0.366	0.286	0.237	0
		Complete	0.623	0.430	0.344	0.294	0.230	0
		Single	0.558	0.269	0.137	0.113	0.054	0

20%	Spearman	Average	0.664	0.485	0.413(*)	0.342	0.288	0
		Ward's	0.704(*)	0.449	0.348	0.318	0.246	1
		Complete	0.684	0.441	0.370	0.261	0.230	0
		Single	0.507	0.366	0.153	0.113	0.040	0
	Kendall's Tau	Average	0.183	-0.009	-0.021	-0.013	0.005	0
		Ward's	0.062	0.066	0.011	0.010	0.002	0
		Complete	-0.244	0.030	0.013	-0.006	-0.002	0
		Single	0.059	-0.026	0.003	0.003	-0.002	0
	Winsorized	Average	0.675	0.473	0.412	0.365(*)	0.284	0
		Ward's	0.671	0.447	0.359	0.306	0.233	0
		Complete	0.621	0.422	0.362	0.290	0.229	0
		Single	0.545	0.170	0.225	0.082	0.043	0
	Freq		1	0	0	0	0	
	Pearson	Average	0.714(*)	0.492	0.419	0.363(*)	0.281	1
		Ward's	0.653	0.453	0.367	0.290	0.239	0
		Complete	0.660	0.424	0.361	0.287	0.212	0
		Single	0.591	0.259	0.181	0.086	0.050	0
	Spearman	Average	0.713	0.493(*)	0.420(*)	0.350	0.282	1
		Ward's	0.641	0.462	0.384	0.308	0.238	0
		Complete	0.640	0.477	0.355	0.307	0.232	0
		Single	0.620	0.247	0.157	0.104	0.055	0
	Kendall's Tau	Average	0.095	0.030	-0.030	0.002	-0.002	0
		Ward's	-0.181	-0.024	0.003	0.006	0.005	0
		Complete	0.258	-0.038	-0.006	0.005	0.001	0
		Single	-0.076	-0.022	-0.017	0.008	-0.004	0
	Winsorized	Average	0.674	0.482	0.414	0.351	0.286	0
		Ward's	0.618	0.445	0.358	0.306	0.237	0
		Complete	0.638	0.427	0.359	0.285	0.214	0
		Single	0.586	0.261	0.161	0.088	0.050	0
	Freq		2	0	0	0	0	
30%	Pearson	Average	0.705(*)	0.499(*)	0.427	0.368(*)	0.286	1
		Ward's	0.658	0.406	0.383	0.290	0.232	0
		Complete	0.593	0.414	0.365	0.266	0.228	0
		Single	0.642	0.212	0.198	0.138	0.055	0
	Spearman	Average	0.694	0.486	0.432(*)	0.354	0.285	0
		Ward's	0.610	0.430	0.372	0.309	0.242	0
		Complete	0.621	0.436	0.353	0.292	0.231	0
		Single	0.595	0.173	0.202	0.095	0.048	0
	Kendall's Tau	Average	0.058	-0.024	-0.008	0.002	-0.002	0
		Ward's	-0.083	0.008	0.013	0.004	0.001	0
		Complete	-0.050	-0.009	-0.003	-0.010	0.002	0
		Single	-0.016	-0.005	0.006	0.003	0.002	0
	Winsorized	Average	0.688	0.493	0.417	0.344	0.282	0
		Ward's	0.608	0.451	0.381	0.308	0.245	0

Complete	0.674	0.430	0.374	0.292	0.239	0
Single	0.565	0.263	0.114	0.107	0.045	0
Freq	1	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.16 shows the results of CCC for 16 clustering algorithms under normal distributions, extreme contaminated observations with 10%, 20% and 30% of outliers injected when $p = 20$. At 10% contaminated observations, only Spearman had CCC above 0.7 at $n = 10$ with Ward's linkage. As for the linkage methods, average and Ward's were well performed. But, Ward's excelled than average when $n = 10$ for Pearson and Spearman distances. Complete came third while single linkage stayed the last. When $n = 10$, $n = 50$, $n = 100$, and $n = 200$, the dominant algorithm were *Spearman-Ward's*, *Pearson-average*, *Spearman-average*, *Winsorized-average*, respectively. There was no appropriate algorithm when $n = 500$ as the coefficients were below 0.3. Therefore, it can be inferred that the appropriate clustering algorithm was the combination of *Spearman-Ward's* as it had the largest coefficients among others.

At 20% contaminated observations, Pearson and Spearman had the same one frequency of CCC that was more than 0.7. Winsorized and Kendall's Tau had none. As for the linkage methods, average functioned the best. Ward's and complete functioned at the same level as they were having close coefficients. Single linkage did not functioned effectively. When $n = 10$, and 200, *Pearson-average* had the largest coefficient while when $n = 50$ and 100, the dominant algorithm was *Spearman-average*. Therefore, it can be concluded that *Pearson-average* algorithm was the appropriate algorithm when there was 20% of outliers when $p = 20$ as it had the highest coefficients specifically at $n = 10$ rather than *Spearman-average* even though they had the same frequency of having largest coefficients.

At 30% contaminated observations, only Pearson had one CCC value above 0.7 at average linkage and $n = 10$. Other distances had none. In terms of linkage methods, average remained the first well-performed linkage. Ward's and complete performed equivalently. By comparing Ward's and complete, it showed that Ward's perform well by having higher coefficients than complete, especially when sample size was more than 100. When $n = 10$, 50 and 200, *Pearson-average* was the dominant algorithm. When $n = 100$, the prevalent algorithm was *Spearman-average*. When $n = 500$, there was no proper algorithm. As for the conclusion, *Pearson-average* was the appropriate algorithm as it had the most frequency of having largest coefficients.

Table 4.17

Average value of CCC for 16 clustering algorithms under normal distribution with different percentage of outliers for extreme ($\mu=5$) contaminated observations when the number of variables, $p = 30$.

Contaminated outlier (%)	Correlation-based distance measures	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
			$p = 30$					
			Samples size, n					
			$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
10%	Pearson	Average	0.654	0.472	0.394	0.329	0.255	0
		Ward's	0.679	0.425	0.341	0.273	0.211	0
		Complete	0.668	0.414	0.309	0.246	0.199	0
		Single	0.583	0.241	0.174	0.096	0.048	0
	Spearman	Average	0.695(*)	0.503(*)	0.400(*)	0.339(*)	0.261	0
		Ward's	0.619	0.410	0.336	0.283	0.203	0
		Complete	0.650	0.381	0.293	0.255	0.202	0
		Single	0.605	0.241	0.152	0.077	0.042	0
	Kendall's Tau	Average	0.270	0.035	0.005	-0.015	-0.002	0
		Ward's	0.025	-0.019	0.023	-0.002	-0.002	0
		Complete	-0.248	0.020	0.012	0.002	0.004	0
		Single	-0.202	0.001	0.009	0.006	-0.006	0
	Winsorized	Average	0.688	0.500	0.396	0.323	0.252	0
		Ward's	0.657	0.401	0.317	0.280	0.206	0
		Complete	0.634	0.384	0.304	0.274	0.196	0
		Single	0.544	0.297	0.125	0.066	0.045	0
Freq			0	0	0	0	0	

20%	Pearson	Average	0.674	0.454	0.387	0.325	0.259	0
		Ward's	0.620	0.422	0.343	0.280	0.208	0
		Complete	0.626	0.408	0.305	0.252	0.205	0
		Single	0.514	0.210	0.106	0.077	0.044	0
	Spearman	Average	0.670	0.468(*)	0.399(*)	0.330(*)	0.260	0
		Ward's	0.687	0.403	0.329	0.279	0.214	0
		Complete	0.679	0.412	0.335	0.227	0.204	0
		Single	0.529	0.313	0.164	0.100	0.045	0
	Kendall's Tau	Average	-0.031	0.005	-0.016	0.010	0.003	0
		Ward's	0.004	0.019	0.008	-0.006	0.003	0
		Complete	-0.108	0.018	0.006	0.007	0.002	0
		Single	-0.146	0.006	0.025	-0.007	-0.002	0
	Winsorized	Average	0.689(*)	0.447	0.388	0.329	0.259	0
		Ward's	0.627	0.412	0.358	0.273	0.203	0
		Complete	0.639	0.422	0.356	0.270	0.192	0
		Single	0.579	0.321	0.130	0.100	0.043	0
Freq			0	0	0	0	0	
30%	Pearson	Average	0.680	0.467(*)	0.385	0.324	0.259	0
		Ward's	0.612	0.381	0.328	0.260	0.206	0
		Complete	0.632	0.401	0.335	0.245	0.197	0
		Single	0.502	0.187	0.136	0.117	0.051	0
	Spearman	Average	0.655	0.450	0.397(*)	0.330	0.261	0
		Ward's	0.637	0.410	0.337	0.281	0.215	0
		Complete	0.660	0.415	0.335	0.256	0.210	1
		Single	0.519	0.277	0.167	0.098	0.043	0
	Kendall's Tau	Average	-0.042	-0.035	-0.006	0.017	0.003	0
		Ward's	0.006	0.039	-0.007	0.005	-0.002	0
		Complete	-0.003	-0.002	-0.008	0.007	0.002	0
		Single	0.065	0.006	-0.002	0.011	0.003	0
	Winsorized	Average	0.675	0.462	0.383	0.332(*)	0.259	0
		Ward's	0.699(*)	0.430	0.347	0.274	0.208	1
		Complete	0.677	0.378	0.294	0.241	0.190	0
		Single	0.273	0.273	0.144	0.087	0.045	0
Freq			1	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Lastly, when $p = 30$, overall, under this number of variable, all the CCC values was below 0.7 for any conditions. Thus, in this section, the coefficient between 0.5 to 0.69 was used to determine the dominant distance.

As for 10% of injected outliers, the highest frequency of CCC for 0.5 and 0.69 coefficients were dominated by Spearman, followed by Winsorized and Pearson. Kendall's Tau was the worst at any conditions. In terms of linkage methods, average performed better than others. Then, followed by Ward's and complete. Single linkage was the weakest. In addition, it can be concluded that *Spearman-average* was the appropriate clustering algorithm when there was 10% of outliers as it had the largest coefficients at all sample sizes.

As for the 20% contaminated observations, overall, based on distances, Spearman performed better than Pearson and Winsorized as it had more frequency on the highest value of CCC for different sample sizes. Kendall's Tau stayed the worst. As for the linkage methods, average was the best. Ward's and complete performed equally while single linkage was the worst. When $n = 10$, *Winsorized-average* had the largest coefficient. When $n = 50$ and above, the dominant algorithm was *Spearman-average*. Therefore, it can be concluded that the appropriate clustering algorithm was the pairing of *Spearman-average*.

At 30% contaminated observations, Pearson and Spearman had the same highest frequency of CCC of 0.5 to 0.69. Winsorized came third and Kendall's Tau had none. In terms of linkage methods, average was the best while, Ward's and complete were equivalently functioned well. By comparing Ward's and complete, complete performed well when combined with Pearson at sample size, $n = 10$ to 100. When $n = 10, n = 50, n = 100$ and $n = 200$, the dominant algorithms are *Winsorized-Ward's*, *Pearson-average*, *Spearman-average* and *Winsorized-average*, respectively. When $n = 500$, there was no such appropriate algorithm to be used. As a

conclusion, the appropriate algorithm when there was 30% of outliers was *Winsorized-Ward's* as it has the largest coefficient among others.

Table 4.18

Summary of appropriate clustering algorithms under normal distribution, mild and extreme contaminated observations with 10%, 20% and 30% of outliers.

Contaminated observation	Outliers (%)	Number of variables, p	Distance-linkage	Range value of CCC
Mild	10%	$p = 3$	<i>Spearman-average</i>	0.775 – 0.873
		$p = 5$	<i>Winsorized-average</i>	0.524 – 0.734
		$p = 10$	<i>Spearman-average</i>	0.385 – 0.729
		$p = 20$	<i>Pearson-average</i>	0.353 – 0.694
		$p = 30$	<i>Winsorized-average</i>	0.332 – 0.674
	20%	$p = 3$	<i>Spearman-average</i>	0.777 – 0.889
		$p = 5$	<i>Spearman-average</i>	0.551 – 0.800
		$p = 10$	<i>Spearman-average</i>	0.381 – 0.786
		$p = 20$	<i>Winsorized-average</i>	0.354 – 0.672
		$p = 30$	<i>Pearson-average</i>	0.329 – 0.636
	30%	$p = 3$	<i>Spearman-average</i>	0.771 – 0.831
		$p = 5$	<i>Spearman-average</i>	0.564 – 0.799
		$p = 10$	<i>Spearman-average</i>	0.357 – 0.778
		$p = 20$	<i>Spearman-average</i>	0.355 – 0.664
		$p = 30$	<i>Spearman-average</i>	0.330 – 0.669
Extreme	10%	$p = 3$	<i>Spearman-average</i>	0.776 – 0.857
		$p = 5$	<i>Spearman-average</i>	0.553 – 0.788
		$p = 10$	<i>Pearson-average</i>	0.364 – 0.793
		$p = 20$	<i>Spearman-Ward's</i>	0.318 – 0.704
		$p = 30$	<i>Spearman-average</i>	0.339 – 0.695

20%	$p = 3$	<i>Spearman-average</i>	0.778 – 0.845
	$p = 5$	<i>Spearman-average</i>	0.546 – 0.788
	$p = 10$	<i>Spearman-average</i>	0.372 – 0.718
	$p = 20$	<i>Pearson-average</i>	0.363 – 0.714
	$p = 30$	<i>Spearman-average</i>	0.330 – 0.670
30%	$p = 3$	<i>Spearman-average</i>	0.776 – 0.820
	$p = 5$	<i>Spearman-average</i>	0.560 – 0.812
	$p = 10$	<i>Winsorized-average</i>	0.355 – 0.806
	$p = 20$	<i>Pearson-average</i>	0.368 – 0.705
	$p = 30$	<i>Winsorized-Ward's</i>	0.347 – 0.699

Based on Table 4.8 until Table 4.17, the list of appropriate clustering algorithm according to the number of variables under mild and extreme contamination was summarized into Table 4.18. Based on the range value of CCC, it can be inferred that *Spearman-average* dominated the others as it functioned very effectively for almost all numbers of variables. The decreasing value was shown by the coefficient of cophenetic correlation when p increases.

4.2.3 Clustering Algorithm under Non Normal Gamma Distribution.

In this section, the simulation results under non normal Gamma distribution was presented. The analysis had been done for $p = 3, 5, 10, 20, 30$ and $n = 10, 50, 100, 200, 500$. The statistical performance of clustering is presented in the following tables.

Table 4.19

Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 3$.

Correlation-based distance measure	Clustering method (Linkage)	Cophenetic correlation coefficient					
		$p = 3$					
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	Freq
Pearson	Average	0.888(*)	0.754	0.756	0.750	0.755	5
	Ward's	0.826	0.758	0.757	0.760	0.733	5
	Complete	0.830	0.715	0.703	0.691	0.661	3
	Single	0.784	0.663	0.655	0.692	0.657	1
Spearman	Average	0.847	0.775	0.779	0.799	0.779	5
	Ward's	0.754	0.776	0.767	0.757	0.756	5
	Complete	0.820	0.764	0.784	0.776	0.770	5
	Single	0.634	0.632	0.629	0.632	0.632	0
Kendall's Tau	Average	0.114	0.030	-0.016	0.013	0.003	0
	Ward's	-0.209	0.036	-0.009	0.003	0.004	0
	Complete	0.238	0.058	-0.014	-0.011	0.003	0
	Single	-0.046	-0.032	-0.024	0.008	-0.003	0
Winsorized	Average	0.863	0.839	0.815	0.877	0.857	5
	Ward's	0.877	0.866(*)	0.817	0.879(*)	0.832	5
	Complete	0.842	0.822	0.845(*)	0.813	0.878(*)	5
	Single	0.812	0.834	0.841	0.839	0.822	5
Freq		11	10	10	9	9	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

In real practice, the data is not always normally distributed. Thus, as for non-normal distribution, Gamma distribution had been chosen in this study. Table 4.19 displayed the average value of CCC for the 16 clustering algorithms constructed under non-normal Gamma distributions when $p = 3$. On the whole, Winsorized had the highest frequency of CCC above 0.7 with Spearman came second, and followed by Pearson. This was shown by the numbers stated at the right-last column of the table. Kendall's Tau did not perform well even under this distribution. In terms of linkage method, average, Ward's method and complete linkage were equally well-performed. All of these three linkages were having close coefficients. For Pearson, Spearman and Winsorized, all the coefficients decreased when sample size increased which was

shown by the changes from 11 to 10 and 9. When $n = 10$, the prevalent algorithm was *Pearson-average*. When $n = 50$ and 200, the dominant algorithm was *Winsorized-Ward's*. When $n = 100$ and 500, the dominant algorithm was *Winsorized-complete*. In summary, the appropriate clustering algorithms when $p = 3$ under Gamma distributions was *Winsorized-Ward's* as it had much higher coefficients than *Winsorized-complete* even though they had the same frequency of having largest coefficients in different sample sizes.

Table 4.20

Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 5$.

Correlation-based distance measure	Clustering method (Linkage)	Cophenetic correlation coefficient					
		$p = 5$					Freq
		Samples size, n					
		$n = 10$	$n = 5$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.776	0.652	0.605	0.586	0.571	1
	Ward's	0.726	0.677	0.570	0.548	0.522	1
	Complete	0.730	0.612	0.589	0.550	0.475	1
	Single	0.682	0.486	0.333	0.288	0.087	0
Spearman	Average	0.764	0.683	0.584	0.581	0.564	1
	Ward's	0.741	0.597	0.580	0.528	0.532	1
	Complete	0.728	0.601	0.552	0.551	0.513	1
	Single	0.669	0.540	0.120	0.174	0.184	0
Kendall's Tau	Average	-0.220	0.025	0.001	-0.003	0.002	0
	Ward's	-0.200	-0.088	0.001	0.002	0.003	0
	Complete	-0.166	0.003	0.001	0.010	-0.001	0
	Single	-0.177	0.054	-0.009	-0.008	-0.003	0
Winsorized	Average	0.808(*)	0.769(*)	0.789(*)	0.736	0.736	5
	Ward's	0.770	0.712	0.706	0.753(*)	0.744	5
	Complete	0.766	0.703	0.728	0.748	0.779(*)	5
	Single	0.719	0.634	0.612	0.626	0.605	1
Freq		10	3	3	3	3	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Table 4.20 presents the average value of CCC for the 16 clustering algorithms constructed under Gamma distribution when $p = 5$. Overall, Winsorized had the highest frequency of having coefficients above 0.7, while Pearson and Spearman had

the same frequency. Kendall's Tau did not perform well. In terms of linkage, average, Ward's and complete are equivalently well-functioned. Single linkage did not function well. For Pearson, Spearman and Winsorized, all the coefficients decreased when sample size increased which was shown by the changes of frequency from 10 to 3. When $n = 10$ to 100, the prevalent algorithm was *Winsorized-average*. When $n = 200$ and $n = 500$, the largest coefficients were *Winsorized-Ward's* and *Winsorized-complete*. Therefore, the appropriate clustering algorithms when $p = 5$ was *Winsorized-average*.

Table 4.21

Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 10$.

		Cophenetic correlation coefficient					
Correlation-based distance measure	Clustering method (Linkage)	$p = 10$					Freq
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.767	0.572	0.513	0.483	0.419	1
	Ward's	0.717	0.504	0.466	0.438	0.378	1
	Complete	0.690	0.522	0.462	0.407	0.365	0
	Single	0.630	0.336	0.301	0.098	0.053	0
Spearman	Average	0.731	0.550	0.493	0.444	0.373	1
	Ward's	0.681	0.495	0.470	0.384	0.324	0
	Complete	0.655	0.496	0.445	0.395	0.327	0
	Single	0.648	0.274	0.153	0.183	0.069	0
Kendall's Tau	Average	0.175	0.017	0.004	-0.006	0.003	0
	Ward's	-0.151	-0.013	0.004	0.007	0.002	0
	Complete	-0.122	-0.039	0.036	0.008	-0.002	0
	Single	-0.122	0.022	0.035	-0.008	-0.003	0
Winsorized	Average	0.772(*)	0.679(*)	0.694(*)	0.643(*)	0.619(*)	1
	Ward's	0.662	0.664	0.628	0.576	0.525	0
	Complete	0.692	0.629	0.600	0.581	0.544	0
	Single	0.660	0.566	0.526	0.510	0.410	0
Freq		4	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

When $p = 10$, Pearson had the highest frequency for the coefficients of CCC above 0.7. Spearman and Winsorized had the same frequency while Kendall's Tau had none. As for the linkage, average was the best while Ward's and complete methods functioned on the same level and exhibited close values of CCC with each other. Single linkage didn't work well except when $n = 10$ and when combined with Winsorized as the distance. In addition, under this number of variables, the CCC values had start to deteriorate below 0.7 for all condition as sample size increased. Next, when $n = 10$ to 500, the most dominant algorithm was *Winsorized-average*. Hence, *Winsorized-average* was the appropriate clustering algorithm.

Table 4.22

Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 20$.

Correlation -based distance measure	Clustering method (Linkage)	Cophenetic correlation coefficient					
		$p = 20$					
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	Freq
Pearson	Average	0.796(*)	0.513	0.438	0.381	0.313	1
	Ward's	0.666	0.481	0.380	0.325	0.276	0
	Complete	0.675	0.414	0.394	0.324	0.270	0
	Single	0.588	0.272	0.151	0.131	0.056	0
Spearman	Average	0.681	0.475	0.435	0.341	0.291	0
	Ward's	0.646	0.456	0.374	0.303	0.239	0
	Complete	0.608	0.405	0.346	0.297	0.227	0
	Single	0.654	0.325	0.146	0.102	0.053	0
Kendall's Tau	Average	-0.101	0.019	-0.014	-0.006	0.001	0
	Ward's	0.170	-0.011	-0.006	0.006	0.002	0
	Complete	-0.031	0.018	0.004	-0.009	-0.003	0
	Single	-0.236	0.049	-0.007	-0.005	-0.002	0
Winsorized	Average	0.665	0.572(*)	0.550(*)	0.558(*)	0.533(*)	0
	Ward's	0.612	0.509	0.484	0.507	0.465	0
	Complete	0.629	0.548	0.485	0.437	0.453	0
	Single	0.473	0.383	0.306	0.351	0.392	0
Freq		1	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Next, when $p = 20$, overall, all the CCC values was below 0.7 for any condition as sample size increased except for *Pearson-average* at $n = 10$. On the whole, the highest frequency of CCC for 0.5 and 0.69 coefficients were dominated by Winsorized, followed by Pearson and Spearman. Kendall's Tau did not function well as usual. In terms of linkage method, average remained as the best linkage. Ward's and complete were equally well performed while single linkage was the worst. When $n = 10$, *Pearson-average* had the largest coefficient. When $n = 50$ and above, *Winsorized-average* was the most dominant algorithm. Thus, it can be concluded that the most appropriate algorithm when $p = 20$ under Gamma distribution was *Winsorized-average*.

Table 4.23

Average value of CCC for 16 clustering algorithms under non-normal Gamma distribution when the number of variables, $p = 30$.

Correlation-based distance measure	Clustering method (Linkage)	Cophenetic correlation coefficient					Freq
		$p = 30$					
		Samples size, n					
		$n = 10$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	
Pearson	Average	0.681	0.493	0.436	0.347	0.286	0
	Ward's	0.689(*)	0.432	0.377	0.306	0.248	0
	Complete	0.626	0.438	0.352	0.266	0.201	0
	Single	0.600	0.235	0.183	0.084	0.072	0
Spearman	Average	0.678	0.482	0.383	0.333	0.257	0
	Ward's	0.639	0.414	0.335	0.269	0.212	0
	Complete	0.681	0.410	0.306	0.237	0.201	0
	Single	0.541	0.233	0.109	0.103	0.034	0
Kendall's Tau	Average	0.122	0.043	-0.015	-0.005	-0.001	0
	Ward's	0.083	0.041	0.008	-0.005	0.001	0
	Complete	-0.086	-0.010	0.009	-0.002	-0.002	0
	Single	-0.105	-0.018	-0.020	0.001	0.002	0
Winsorized	Average	0.613	0.559(*)	0.519(*)	0.477(*)	0.485(*)	0
	Ward's	0.563	0.497	0.438	0.433	0.411	0
	Complete	0.549	0.476	0.426	0.372	0.384	0
	Single	0.485	0.342	0.219	0.234	0.241	0
Freq		0	0	0	0	0	

Note : CCC that was 0.7 and above (bold), between 0.3 to 0.699 (black), less than 0.3 (blue). The largest coefficient for each sample size (*). Freq = frequency of CCC that was 0.7 and above.

Lastly, when $p = 30$, overall, the highest frequency of CCC for 0.5 and 0.69 was dominated by Winsorized then followed by Pearson and Spearman. Kendall's Tau did not perform well in every sample size and linkage methods as predicted. As for the linkage methods, average outperformed others. Ward's and complete were equally well-performed as they were having close value of CCC. Single was the last. When $n = 10$, *Pearson-Ward's* had the largest coefficient while when $n = 50$ and above, *Winsorized-average* had dominated by having largest coefficients. Thus, as for conclusion, when $p = 30$, under Gamma distributions, the appropriate clustering algorithm was the combination of *Winsorized-average*.

Table 4.24

Summary of appropriate clustering algorithm under non-normal Gamma distribution

Number of variables, p	Distance-linkage	Range value of CCC
$p = 3$	<i>Winsorized-Ward's</i>	0.817 – 0.879
$p = 5$	<i>Winsorized-average</i>	0.736 – 0.808
$p = 10$	<i>Winsorized-average</i>	0.619 – 0.772
$p = 20$	<i>Winsorized-average</i>	0.533 – 0.665
$p = 30$	<i>Winsorized-average.</i>	0.485 – 0.613

Based on Table 4.19 until Table 4.23, we summarized the list of appropriate clustering algorithm according to the number of variables under non-normal Gamma distribution into Table 4.24. Based on the range value of CCC, it can be inferred that *Winsorized-average* dominated the others as it functioned very effectively for almost all numbers of variables, except when $p = 3$. Even under non-normal Gamma distributions, when p increased, the coefficient of cophenetic correlation decreased.

Based on the simulated results discussed in previous sections, the results have been summarized according to the number of variables as in Table 4.26 to Table 4.30. Before that, the indicator that has been used was presented in Table 4.25.

Table 4.25

Indicators for Table 4.26 until Table 4.30

Alphabet	Indicators
p	Number of variables
n	Sample size
L	Linkage method
A	Average
W	Ward's
C	Complete
S	Single
D	Distance
P	Pearson
S	Spearman
K	Kendall's Tau
W	Winsorized

Table 4.26 Summary of simulated results when number of variables, $p = 3$

p	n	L	Normal			Mild Contaminated												Extreme Contaminated												Gamma		
						10%				20%				30%				10%				20%				30%						
			D	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	
3	10	A				X					X			X					X			X							X			
	50					X				X		X						X								X						
	100				X			X						X			X		X		X		X		X							
	200				X				X					X			X		X		X		X		X							
	500				X			X						X			X		X		X		X		X							
	10	W	X																							X						
	50																								X					X		
	100																															
	200																															
	500																															
	10	C																														
	50																															
	100													X																X		
	200								X																							
	500																													X		
	10	S																														
	50																															
	100																															
	200																															
	500																															

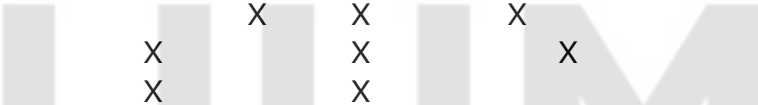


Table 4.28 Summary of simulated results when number of variables, $p = 10$

		Normal					Mild Contaminated										Extreme Contaminated										Gamma														
							10%					20%					30%					10%					20%					30%									
		D	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W			
p	n	L																																							
10	10		X					X				X				X				X						X												X			
	50			X				X					X				X				X					X												X			
	100	A		X			X				X				X				X				X			X												X			
	200					X	X				X				X				X				X				X											X			
	500		X					X				X			X				X				X			X				X								X			
	10																																								
	50																																								
	100	W																																							
	200																																								
	500																																								
	10																																								
	50																																								
	100	C																																							
	200																																								
	500																																								
	10																																								
	50																																								
	100	S																																							
	200																																								
	500																																								

Table 4.29 Summary of simulated results when number of variables, $p = 20$

		Normal					Mild Contaminated										Extreme Contaminated										Gamma												
							10%					20%					30%					10%					20%					30%							
		D	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W					
p	n	L																																					
20	10						X							X	X							X				X				X									
	50						X			X				X		X			X			X			X				X					X					
	100	A				X		X			X							X		X			X			X			X					X					
	200			X				X					X		X					X	X			X			X							X					
	500																																		X				
	10																X																						
	50			X																																			
	100	W																																					
	200																																						
	500																																						
	10		X																																				
	50																																						
	100	C																																					
	200																																						
	500																																						
	10																																						
	50																																						
	100	S																																					
	200																																						
	500																																						

Table 4.30 Summary of simulated results when number of variables, $p = 30$

			Normal					Mild Contaminated												Extreme Contaminated												Gamma		
								10%				20%				30%				10%				20%				30%						
			D	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W	P	S	K	W			
p	n	L																																
30	10	A																																
	50																																	
	100																																	
	200																																	
	500																																	
	10	W																																
	50																																	
	100																																	
	200																																	
	500																																	
	10	C																																
	50																																	
	100																																	
	200																																	
	500																																	
	10	S																																
	50																																	
	100																																	
	200																																	
	500																																	

From the results, it is evident that the suitability of different clustering algorithms depends on the specific conditions of the data or observations. To illustrate, the *Spearman-average* clustering algorithm is the most appropriate choice when dealing with normal and contaminated distributions. It demonstrates exceptional performance across various number of variables and sample sizes, particularly when handling distribution containing outliers. This algorithm had also been applied on the real water quality parameter data as based on the simulation results in which will be discussed in the next sections in this chapter. Furthermore, when dealing with non-normal Gamma distribution, the *Winsorized-average* algorithm emerges as the most suitable, consistently showcasing highest CCC value for a majority number of variables. The *Pearson-average* algorithm also displays substantial CCC values under several data conditions. On top of that, it can be observed that as the number of variables increases, starting when $p = 20$ and the samples size is $n = 500$, under the normal and contaminated distributions, there is no such appropriate algorithm can be used as the value of CCC are below 0.3. This indicates that the clustering algorithm has become less effective when the number of variables and sample size increases.

4.3 Real Data Screening

Before applying the appropriate clustering algorithm on the real water quality parameters dataset, a statistical data screening procedure have been implemented in order the check the normality of the data. The statistical data screening that was being executed was multivariate normality test. The test started by calculating the Mahalanobis distance for the five water quality parameters variables for 57 number of observations by using SPSS software. Then, all of the Mahalanobis distance ($d^2(j)$) values were transferred to Excel sheet and had been sorted from the

smallest to the largest value. Next, the observed cumulative frequency, $(j - 0.5)/n$ and the Chi Square value ($q(j)$) which consist of probability and degree of freedom in its formula were computed as in the Table 4.31.

Table 4.31

Computation results of multivariate normality test

Observations	$d^2(j)$	$(j - 0.5)/n$	$q(j)$	Observations	$d^2(j)$	$(j - 0.5)/n$	$q(j)$
1	0.84552	0.008621	0.52	30	4.42872	0.508621	4.41
2	0.99497	0.025862	0.84	31	4.43473	0.525862	4.54
3	1.26214	0.043103	1.07	32	4.53927	0.543103	4.67
4	1.32865	0.060345	1.25	33	4.73716	0.560345	4.81
5	1.39412	0.077586	1.42	34	4.78036	0.577586	4.95
6	1.43381	0.094828	1.57	35	4.82147	0.594828	5.09
7	1.44127	0.112069	1.71	36	4.85599	0.612069	5.23
8	1.4922	0.12931	1.84	37	4.92661	0.62931	5.39
9	1.49721	0.146552	1.97	38	5.00449	0.646552	5.54
10	1.56586	0.163793	2.09	39	5.07537	0.663793	5.70
11	1.6413	0.181034	2.21	40	5.11237	0.681034	5.87
12	1.68888	0.198276	2.33	41	5.73938	0.698276	6.05
13	1.78338	0.215517	2.45	42	6.00948	0.715517	6.23
14	1.79704	0.232759	2.56	43	6.17971	0.732759	6.42
15	1.7971	0.25000	2.67	44	6.76837	0.75000	6.63
16	1.94141	0.267241	2.79	45	7.32195	0.767241	6.84
17	2.18295	0.284483	2.90	46	7.57844	0.784483	7.07
18	2.55791	0.301724	3.01	47	7.80951	0.801724	7.31
19	2.61753	0.318966	3.12	48	8.26936	0.836207	7.87
20	2.69308	0.336207	3.24	49	8.77479	0.853448	8.18
21	2.8488	0.353448	3.35	50	8.94044	0.87069	8.53
22	2.96797	0.37069	3.46	51	9.74439	0.887931	8.93
23	3.11733	0.387931	3.58	52	10.13194	0.905172	9.38
24	3.50134	0.405172	3.69	53	10.46884	0.922414	9.92
25	3.5288	0.422414	3.81	54	10.83885	0.939655	10.58
26	3.56174	0.439655	3.92	55	12.02629	0.956897	11.45
27	3.5659	0.456897	4.04	56	14.32192	0.974138	12.75
28	3.68857	0.474138	4.17	57	18.33369	0.991379	15.45
29	4.23931	0.491379	4.29				

Then, the values of $d^2(j)$ against $q(j)$ is plotted in the scatter chart as follow.

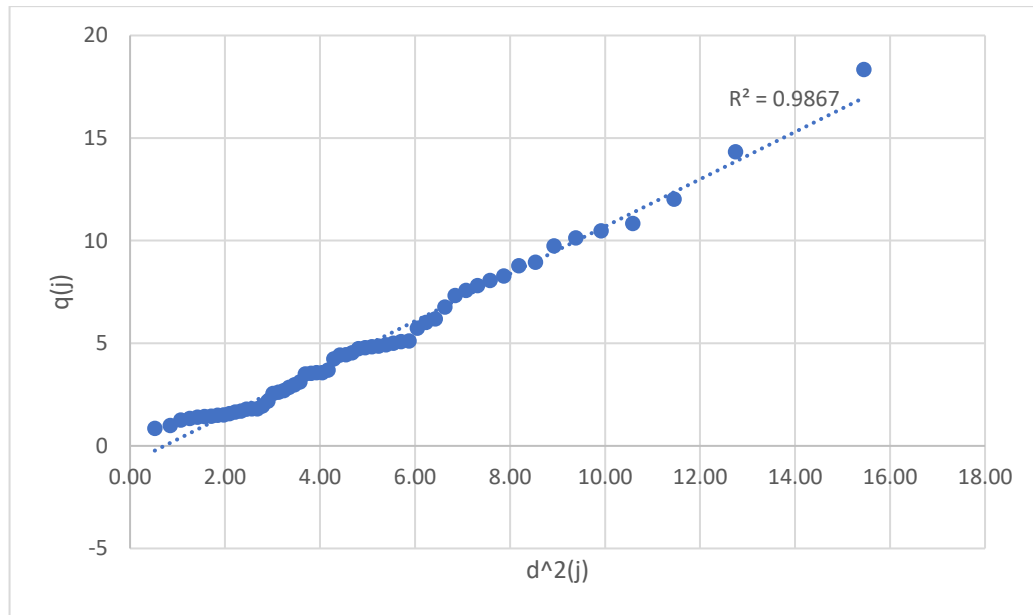


Figure 4.1 Multivariate normality scatter plot

From Figure 4.1, the data was normally distributed ($R^2 = 0.987$) as all the data points in this plot are located approximately along a straight line. All observations value was fitted to the expected distribution value. Thus, it can be concluded that the selected distribution correctly represents the data or in simple words the normality of the water quality data was normal. No outliers were detected in this real data screening. Thus, based on the simulation study, the appropriate clustering algorithm under normal distribution when $p = 5$ and $n = 50$ was *Spearman-average* had been chosen to be implemented on real data. This is because the real data also consists the same $p = 5$ and the $n = 57$.

4.4 Real data analysis using Juru River Basin water quality parameter

Analysing and interpretation of the cluster results is a crucial task. The exploration of the clusters helped us to understand the distinct characteristics and differences of each cluster and also the similarity within the cluster. In this study, 57 observations of five water quality parameters of the Juru River Basin were being used to develop

the clustering model. The analysis of the cluster model involved examining the spider web, correlation matrix, evolution of indices, results by cluster, dendrogram plot, heat map analysis and cluster profile plot.

4.4.1 Descriptive analysis of five water quality parameter

In the first stage, five water quality parameters were explored using spider web to visualize of how well the association between those parameters. Spider web is one of the type of radar charts that serve as a graphical tool used to display multivariate data in a two-dimensional format (V. K. Singh & Tewari, 2022). It consists of a series of axes or spokes radiating from a central point, with each axis representing a different variable or attribute. The data points for each variable were plotted along the corresponding axis, and a line or polygon was formed by connecting these points. The resulting shape provided a visual representation of the relative values or performance of the variables being compared (Kowalczyk & Nizankowski, 2017). In this study, the spider web can be a handy tool for mapping the mean values of the five water quality parameters dataset in which can provide a visual representation on the status of the water quality of Juru River basin throughout the year 2008 to 2017.

Additionally, the mean values of the five parameters can also be used to portray the contamination status of the Juru river basin by classifying them based on Water Quality Index (WQI) under National Water Quality Standards for Malaysia (NWQS) set by DOE. Under NWQS, Water Quality Index (WQI) is developed and serves as the guideline to classify and determine the water quality status of the river based on few parameters. This index as according to Environmental Quality Report by Malaysian Department of Environment in the year 2020 is tabulated as shown in Table 4.32.

Table 4.32

WQI classes and contamination statuses

Parameter (unit) /Class	Class I	Class II	Class III	Class IV	Class V
NH ₃ -NL (mg/l)	< 0.1	0.1 – 0.3	0.3 – 0.9	0.9 – 2.7	> 2.7
BOD (mg/l)	< 1	1- 3	2 - 6	6 – 12	> 12
COD (mg/l)	< 10	10 - 25	25 - 50	50 – 100	> 100
DO (mg/l)	> 7	5 - 7	3 - 5	1 - 3	< 1
pH (mg/l)	> 7.0	6.0 – 7.0	5.0 – 6.0	< 5.0	> 5.0
TSS (mg/l)	< 25	25 - 50	50 – 150	150 - 300	> 300
Contamination status	Not contaminated	Slightly contaminated	Moderately contaminated	Severely contaminated	Extremely contaminated

Based on Table 4.32, there were six parameters chosen and set for the WQI which were DO, BOD, COD, TSS, NH₃-NL and pH. However, in this study, the pH parameter was not being chosen as the values of the data collected have no variation with each other. The pH value was around 6 to 7. Next, the mean values of the five water quality parameters were depicted in the form of spider web chart as shown in Figure 4.2.

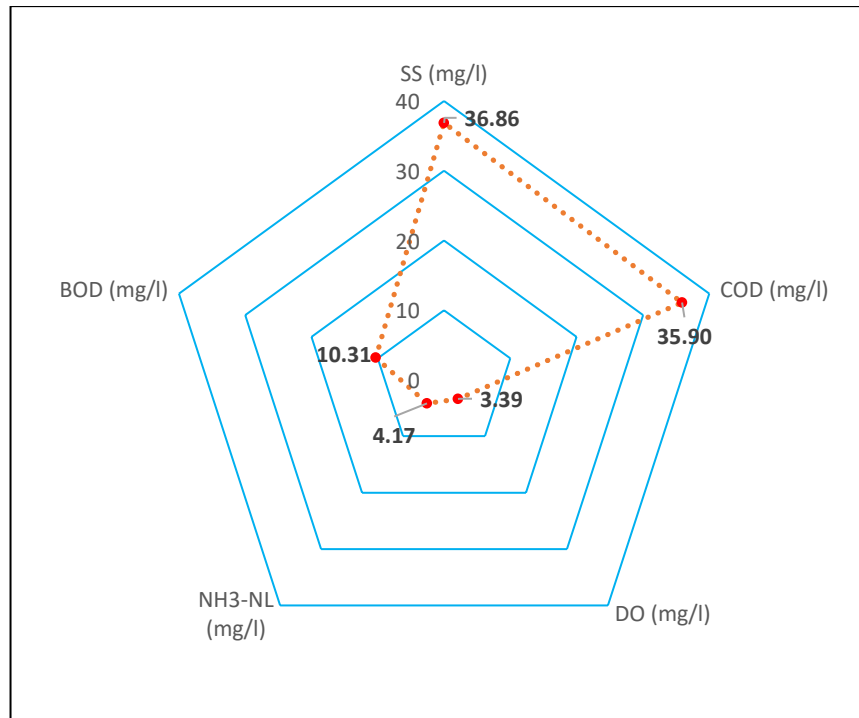


Figure 4.2 Spider web overview of water quality parameters

In Figure 4.2, low mean values were located closest to the centre of the web which was closer to zero. Meanwhile, higher values were positioned farther towards outermost ring (Kowalczyk & Nizankowski, 2017). Based on the spider web, COD (35.90mg/l) and TSS (36.86mg/l) parameters were at the outermost ring. COD were considered to be moderately high concentration and categorized as Class III while TSS was slightly high concentration and fell under Class II according to WQI threshold level for the Malaysian rivers. In contrast, parameter DO (3.39mg/l) that was the closest to the centre of the web indicated that there was low concentration of dissolved oxygen. It is classified under Class III of WQI. Class III of this WQI standard indicated that the water supply is moderately contaminated and extensive treatment was needed. As for the parameters NH₃-NL(4.17mg/l) and BOD (10.31mg/l), their concentration was considered to be severely high and being classified as Class V and Class IV threshold level of WQI. Class V and IV threshold

indicates that the river was severely contaminated and not suitable for human consumptions.

In the next subsections, the result of clustering model was presented based on correlation matrix, clustering size, evolution of indices and in the form of dendrogram. The practical value of cluster analysis was presented in its application to evaluate and interpret water quality data which might be helpful for the purpose of monitoring water resource management.

4.4.2 Correlation Matrix

Correlation matrix is a matrix that portrays the pairwise relationships between a set of variables. It provides a comprehensive view of the relationships between variables of interest. Elements in the matrix represents the correlation coefficient, which indicates the strength and direction of the relationship between two variables. Based on Table 4.27, the darker colour indicated the stronger correlation value where the value is close to one (or negative one). Meanwhile, the brighter colour elements represent the weaker correlation between variables.

Table 4.33

Correlation matrix for five water quality parameters

Parameter	TSS (mg/l)	COD (mg/l)	DO (mg/l)	NH3-NL (mg/l)	BOD (mg/l)
TSS (mg/l)	1	0.389	0.104	0.179	0.227
COD (mg/l)	0.389	1	-0.009	0.310	0.647
DO (mg/l)	0.104	-0.009	1	0.103	0.002
NH3-NL (mg/l)	0.179	0.310	0.103	1	0.191
BOD (mg/l)	0.227	0.647	0.002	0.191	1

From Table 4.33, water quality parameters COD and BOD showed the largest correlation with the positive value of 0.647 while DO and BOD showed the smallest correlation with the positive value 0.002. The second largest correlation was 0.389 between parameters COD and TSS, followed by the correlation value of 0.310 which was between parameters COD and NH₃-NL. This analysis showed that COD were correlated to three parameters in positive direction except for COD and DO (weak negative correlation coefficient, -0.009).

By referring to NH₃-NL and TSS, all of their relationship to the others were within 0.103 and 0.389. It indicated both were not strongly correlated. Lastly, the correlation between DO and the other parameters produced small correlation coefficient. DO have had very low correlations coefficient with the parameter BOD and COD with the values of 0.002 and -0.009, respectively. From this analysis, it can be concluded that all the parameters were interrelated to each other except for DO. However, to discriminate between the cluster, DO was taking into consideration. From literature, it reported that if DO levels were too high, water cannot hold onto other dissolved substances. If DO levels were too low, minerals from riverbeds and runoff will start to dissolve into the water, which affects water quality. This statement showed the important of this parameter to be included in the clustering model.

4.4.3 Clustering Model Analysis

In this section, the clustering model of Juru river basin water quality using *Spearman-average* was presented using five water quality parameters and 57 observations. Table 4.34 presents the evolution of indices to validate the clusters

result that have been produced by using the real water quality data consist of CCC and Calinski & Harabasz index.

Table 4.34

Evolution indices for real data when $p = 5$ and $n = 57$.

Cophenetic correlation (CCC)	0.76			
Number of clusters	2	3	4	5
Calinski & Harabasz index	17.077	20.138	22.511	26.879

The CCC which formula referred to equation (4.5) was used in measuring the goodness of clustering model which reflecting to intergroup dissimilarity. Based on Table 4.34, the CCC was 0.76 which indicated high correlation coefficient. This signified that there was high correlation between the distance matrix and cophenetic distance matrix. In other words, the dendrogram produced from this clustering algorithm had faithfully represent or represent well the pairwise distances between the original observations. Furthermore, Calinski & Harabasz index which formula referred to equation (2.4) was 26.879 which indicated that the optimal number of clusters for the clustering model was five clusters.

From dendrogram in Figure 4.3, the five clusters namely C1 (turquoise), C2 (blue), C3 (green), C4 (red), and C5 (purple) were clearly separated from each other. The index measured the ratio of between-cluster variance to the within-cluster variance. Thus, it implied that the data points within each cluster were similar, and the clusters were distinct from one another, making the clustering solution more reliable and meaningful.

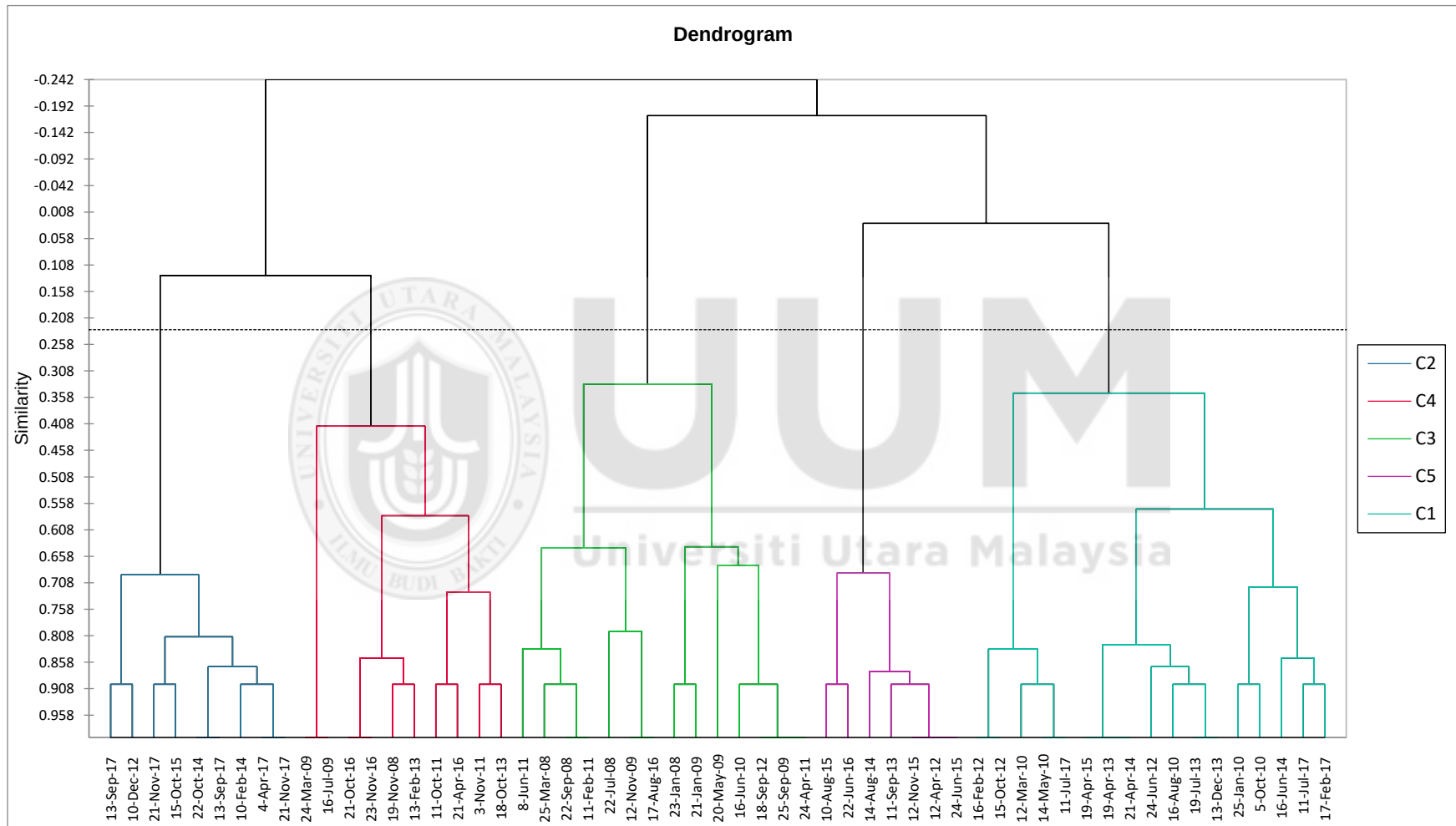


Figure 4.3 Agglomerative hierarchical clustering dendrogram for number of clusters = 5

Five clusters were derived from the clustering model and depicted in the form of dendrogram as shown in Figure 4.3. This dendrogram illustrated how the clustering algorithm works in grouping the observations and subsequently forming subgroups of observations. As evident from the results, the algorithm has successfully clustered all the observations. Additionally, the dotted line represents the truncation which led to five clusters. Next, the five number of clusters resulted from the clustering model was described further in Table 4.35.

Table 4.35

<i>Results by cluster</i>					
Clusters	1	2	3	4	5
Number of observations by cluster	17	9	14	10	7
Within-cluster variance	682.540	114.092	385.660	682.273	223.120
Average distance to centroid	20.481	9.310	16.682	19.670	11.995
Observations in each cluster	11-Jul-17 25-Jan-10 12-Mar-10 14-May-10 16-Aug-10 5-Oct-10 16-Feb-12 24-Jun-12 15-Oct-12 19-Apr-13 19-Jul-13 13-Dec-13 21-Apr-14 16-Jun-14 19-Apr-15 17-Feb-17 11-Jul-17	13-Sep-17 21-Nov-17 10-Dec-12 10-Feb-14 22-Oct-14 15-Oct-15 4-Apr-17 13-Sep-17 21-Nov-17	23-Jan-08 25-Mar-08 22-Jul-08 22-Sep-08 21-Jan-09 20-May-09 25-Sep-09 12-Nov-09 16-Jun-10 11-Feb-11 24-Apr-11 8-Jun-11 18-Sep-12 17-Aug-16	19-Nov-08 24-Mar-09 16-Jul-09 11-Oct-11 3-Nov-11 13-Feb-13 18-Oct-13 21-Apr-16 21-Oct-16 23-Nov-16	12-Apr-12 11-Sep-13 14-Aug-14 24-Jun-15 10-Aug-15 12-Nov-15 22-Jun-16

Table 4.35 presented five clusters that have been derived from the appropriate clustering algorithm. C1 make up the biggest group comprising 17 number of objects in the cluster while C3 was the second biggest cluster containing 14 number of objects. On the other hand, C2 and C4 have close number of objects which is nine and ten in the cluster respectively. Whereas, the smallest cluster was C5 comprising seven objects in the cluster.

When looking at within-cluster variance, C2 has the smallest value, 114.092 which indicated the cluster was the most homogenous cluster among other. Homogenous cluster here means that the similarity between all objects in the cluster was high. The second homogenous cluster was C5 with the value of 223.120. C3 and C4 had the within-cluster variance values of 385.660 and 682.273, respectively. Meanwhile, C1 was the most heterogenous cluster amongst other as it had the highest within-cluster variance with the value of 682.540.

The value of within-cluster variance is related to the average distance to centroid. The higher value of within-cluster variance will have large average distance to centroid, vice versa. It shows that the observations in the cluster lies farther from the centroid of the cluster. Cluster that has a small average distance to centroid is more condense compare to cluster that has a large average distance.

Next, when observing to the average distance to centroid value, C2 remained the lowest value which was 9.310 while the highest value, 20.481 was belongs to C1. C5 had the second lowest average distance to centroid value which was 11.995. Subsequently, C3 was 16.682 and C4 had the average distance to centroid value of 19.670. Clusters that have higher values exhibit greater variability of the observations within the cluster. Thus, it can be deduced that C2 was the most

compact cluster among others while C1 was the loose clusters (observations were too dispersing).

Hence, as a whole, it can be concluded that C2 was the most well-defined, cohesive, and highly separated from other clusters as it has the lowest values of within-cluster variance and average distance to centroid. C5 came second followed by C3 and then C4. C1 came last as it had the highest values of within-cluster variance and larger average distance to centroid which indicated that the data points within the cluster were diverse and dissimilar to each other. High value within-cluster variance indicates that the data points in the cluster were scattered and do not exhibit strong similarity or cohesion. In practice, it is easy to handle the observations that has been clustered under C2 compared to C1. In the next section, the mean centroids for each cluster were depicted in the form of heat map analysis and cluster profile plot.

The heat map analysis, which is perfect for large-scale data visualization, uncovers multiple significant attributes within each cluster, demonstrating their strong correlation. The colour scale used in the analysis shows the attribute importance, with lighter yellow representing low centroid values and darker yellow indicating high centroid values for each parameter.

Table 4.36

List of mean centroids for each cluster from heat map analysis

	TSS (mg/l)	COD (mg/l)	DO (mg/l)	NH3-NL (mg/l)	BOD (mg/l)
Cluster 1 (C1)	33.882	43.024	2.405	4.790	14.765
Cluster 2 (C2)	17.444	22.422	4.256	4.868	8.000
Cluster 3 (C3)	52.286	42.500	2.748	4.611	8.286
Cluster 4 (C4)	51.100	27.700	4.008	3.065	6.300
Cluster 5 (C5)	22.286	37.571	4.687	2.235	13.000

Cell colour					
Concentration levels indicator	Very Low	Low	Moderate	Severely high	Extremely high

Table 4.36 displayed the list of mean centroids for each cluster in the form of heat map analysis. The colour scale used in this analysis was based on the parameters' concentration levels and also the WQI classes as tabulated in Table 4.31. WQI was being referred to as a guideline for the indicators because the range of mean centroids value for each parameter has different interpretation from one another. Note that the low value for parameter DO was being classified under Class IV and V indicating that the river might be severely contaminated and extremely contaminated while high value of DO indicate that the river was not or slightly contaminated as categorized in Class I and II. This is because DO serves as an indicator of water body pollution, where low value of DO corresponds to a high level of contamination in the water source and vice versa. Overall, the results showed that Class III, IV and V dominated the parameter level. From Table 4.36, the information was then transferred into profile plot (Figure 4.4) to visualize the performance of each cluster based on five different parameters. The observations for each cluster which is the

time period of the data being collected were referred based on Table 4.29 and also being used to elaborate further on the behaviour of these clusters.

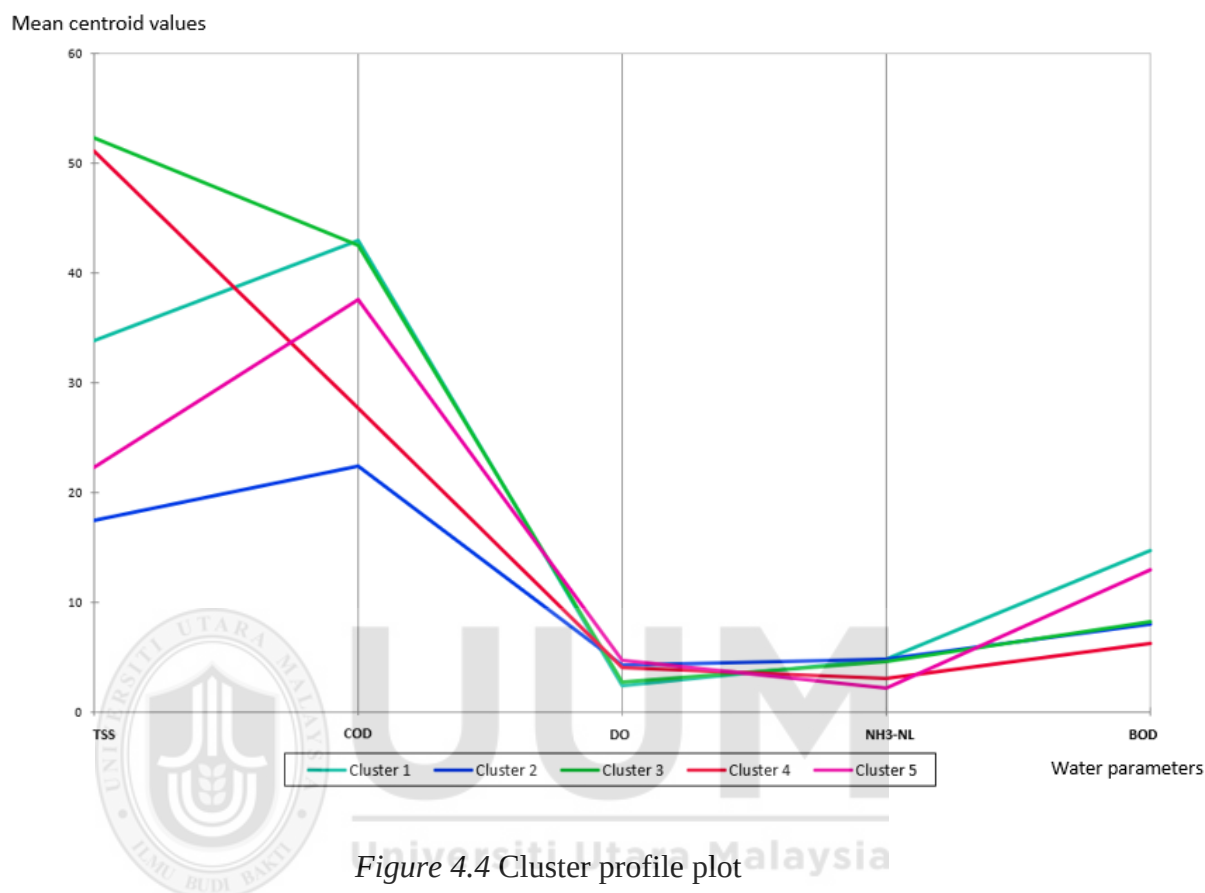


Figure 4.4 Cluster profile plot

Figure 4.4 showed the cluster profile plot with y-axis indicated the mean centroid values and x-axis was the five water parameters. Based on the mean centroid values in each cluster in Table 4.30, C1 had extremely high centroid values for parameters NH3-NL and BOD with the values of 4.790mg/l and 14.765mg/l, respectively. Additionally, it can be observed that the centroid values for parameters BOD (14.765mg/l) and COD (43.024mg/l) in C1 was the highest among other clusters whereas parameter TSS (33.882mg/l) was categorized as low value. The centroid value of parameter DO(2.405mg/l) in C1 was the lowest among other clusters and categorized as Class IV based on WQI. From Figure 4.4, C1(turquoise colour line) had the highest concentration of COD and BOD and also the lowest concentration of

DO parameters among other clusters. Hence, it was concluded that the condition of Juru river basin during the, days, months and years in Cluster 1 is considered to be most contaminated river during that time. This could be due to discharge of extensive organic, industrial sewage and domestic sewage, and agricultural runoff that carries fertilizer, pesticides into the river which lead to excessively high COD and BOD levels. In addition, the low level of DO also shows that the river was highly polluted (Toriman et al., 2011). This is because there was a depletion of oxygen when there was substantial amount of organic and chemical pollution. Besides, the mean centroid value of the NH₃-NL parameter in C1 (4.790 mg/l), falls under Class V according to the WQI threshold which indicated that the river has been extremely contaminated. Furthermore, the data for the months falling within this cluster were February, April and June, which were in the middle of the year. Considering the days when the data was being collected, most of the days were during the early and the end of weekdays which were Monday and Tuesday. Thus, the months and the days in C1 was considered to be active periods which human activities were perceived to have just begun and became active.

Next, the river was considered the second most polluted during the periods in C3. By referring to C3, it had the highest TSS centroid value amid others with the value of 52.286 mg/l. Parameters COD, DO, NH₃-NL and BOD were classified as moderate, low, extremely high and severely high with the values of 42.500mg/l, 2.748mg/l, 4.611mg/l and 8.286mg/l, respectively. In the case C3 (green color line), it exhibited the highest concentration of TSS, followed by COD and NH₃-NL parameters. BOD was considered to be at a severely high level of concentration. In contrast, DO parameter in C3 was considered to be low concentration and being classified under Class IV based on WQI threshold. Therefore, it is inferred that the condition of the

Juru river basin in Cluster 3 was the second most contaminated after Cluster 1 as the DO value is high and COD, NH₃-NL, BOD values is lesser in Cluster 3 than those in Cluster 1. The high levels of TSS, COD and NH₃-NL, resulting from runoff originating from industrial and urban areas in Juru, agriculture, and livestock waste, indicate elevated concentrations of suspended solids and ammonia. Moreover, when the concentration of TSS parameter was high, it showed that there were many contaminants of suspended solids in the river. These levels can affect light penetration and impede oxygen exchange at the water's surface. As for the month and day when the data was being collected, most of the months in C3 was January and day was Wednesday which was earlier of the year and the middle of the weekdays, whereby the human activities are moderately active.

The river that perceived as the third most polluted during years and months was in C4. It has moderate values for parameters TSS, COD and DO with the values of 51.100mg/l, 27.700mg/l and 4.008mg/l, respectively. Parameter NH₃-NL and BOD were classified as extremely high and severely high with their values were 3.065mg/l and 6.300mg/l. As for C4 (red color line) which had a very different plot trend from the others had the lowest concentration of BOD parameter than other clusters. The concentration of COD, DO and TSS were considered to be at a moderate level of concentration. Additionally, it can be observed that the concentration of the TSS parameter was the second highest after C3, while the DO parameter was lower than in C2 and C5. Besides, most of the days in C4 was in Thursday and Friday which reach towards the weekend. Thus, the human activities might had been slowed down. Therefore, it can be deduced that the Juru river basin was moderately polluted during the days, months and years in C4, ranking as the third most polluted after C3.

The river was regarded to be the fourth polluted or second least polluted during the period in C5 (purple colour line). The TSS (22.286mg/l), COD (37.571mg/l) and BOD (13.000mg/l) parameters were classified as very low, moderate and extremely high, separately. Besides, C5 had the highest mean centroids of parameter DO (4.687mg/l) and the lowest value of NH₃-NL (2.235mg/l) amongst the other clusters. Even though C5 has the highest concentration of DO and the lowest concentration of NH₃-NL, it was regarded as the second least polluted due to its high concentrations of TSS, COD, and BOD compared to C2. These indicators led to the conclusion that the condition of the Juru river basin during the times in C5 was considered to be slightly more contaminated than in C2. Besides, COD and BOD were closely connected to the decomposition of organic components in the waterbody. Nonetheless, COD involves the breakdown of the organic and non-organic elements chemically (Hashim et al., 2021). Thus, high level of COD and BOD parameters concentrations shows that the river was polluted. Besides, the data collection days in these clusters predominantly fall on weekdays and middle of month when human and industrial activities are slightly active.

Lastly, Juru river basin was the least polluted during the periods in C2. It has had the highest mean centroid value of parameter NH₃-NL (4.868mg/l) and the lowest value of parameter TSS (17.444mg/l) and COD (22.422mg/l) among other clusters. The other two parameters, DO (4.256mg/l) and BOD (8.000mg/l) can be classified as moderate and severely high value. Being compared the DO levels in C2 and C4, DO in C2 was higher than in C4. The high level of DO indicate that there was high dissolved oxygen in the water that was crucial for the overall health of the river ecosystem. In addition, the levels of TSS, COD, and BOD parameters in C2 were observed to be lower than those in C5. Thus, it can be concluded that Juru river basin

during the months and years in C2 was considered to be the least contaminated. Additionally, the river was deemed to be less polluted, possibly due to the timing of data collection. The months of September, October and November were considered to be towards the year's end. Consequently, industrial activities might be slowing down, potentially resulting in reduced discharge of sewage and other pollutants.

To sum up, based on the parameter and WQI threshold levels, it can be concluded that Juru river basin was the most polluted during the time period in Cluster 1, followed by Cluster 3, Cluster 4, Cluster 5 and the least polluted in Cluster 2. This was because during the period in C1, the river had the highest concentration levels of COD, BOD and the lowest level of DO parameter, followed by C3. Next, C4 was considered to be the third as it had moderate concentration levels for parameters TSS, COD and DO. C5 fell the fourth polluted or second least polluted as it had higher concentration levels of TSS, COD and NH₃-NL compared to C2, although the concentration of DO is higher than C2. C2 was the least polluted due to its lowest TSS and COD levels, and high DO levels which indicate that the river had low decomposition of organic and inorganic elements and a low content of suspended solids.

CHAPTER FIVE

CONCLUSIONS

5.1 Conclusions

The results of clustering rely on the choice and pairing of distance measure and linkage method used in its algorithm. Hence, it is crucial to explore which clustering algorithm(s) excel in comparison to others concerning different distance measures and linkage method. Additionally, the correlation-based distance can also be impacted by outliers, in which will resulting in a distorted interpretation of the clustering outcomes. Therefore, this study's main objective was, to investigate the statistical performance of 16 clustering algorithms constructed from the combination of four correlation-based distance measures (PCC, Spearman rank correlation, Kendall's Tau and Winsorized correlation) and four linkage methods (average linkage, Ward's method, complete linkage and single linkage) using CCC via simulation study under different distributions. The sub-objectives of this study were four which were first, to measure the 16 clustering algorithms performance that were formed by calculating the average value of CCC and second, to compare and determined the appropriate correlation-based distances based on CCC when there were outliers in data under several simulated condition. Thirdly, the study aimed to propose a table of summary as a guideline for choosing appropriate clustering algorithm to the researcher and lastly to verify and demonstrate the appropriate clustering model for Juru River Basin water quality real dataset.

From the simulated results, on a whole, it can be deduced that the suitability of different clustering algorithms depends on the specific conditions of the data or observations. Based on the average value of CCC for the 16 clustering algorithms constructed, it can be seen that *Spearman-average* clustering algorithm was the most

appropriate choice when dealing with normal and contaminated distributions while *Winsorized-average* algorithm emerges as the most suitable when dealing with non-normal Gamma distribution. The *Pearson-average* algorithm also showed significant CCC values under several data conditions. Besides, it can be observed that as the number of variables increases, starting from $p = 20$ with a sample size of $n = 500$, no suitable algorithm can be identified under both normal and contaminated distributions, as the CCC values fell below 0.3. This showed that the clustering algorithm became less effective as the number of variables and sample size increased.

In order to compare and determined the appropriate correlation-based distances based on CCC when there were outliers in data, the data had been simulated under several condition which involved normal distribution with mild and extreme level of data contamination injected with 10%, 20% and 30% percentage of outlier for each level. Based on the findings, *Spearman-average* had appeared as the most appropriate algorithm when there were outliers as it consistently showcasing highest CCC value for a majority number of variables and sample sizes.

Next, for the purpose of summarizing the results of statistical performance in a convenient way to the researcher, the table of summary as the guideline for choosing appropriate clustering algorithm had been proposed and presented in Chapter 4. This table of summary had taken into account factors such as the number of variables, sample size, percentage of outliers, and the type of data distributions which provide valuable insights for researchers in selecting the appropriate distance and linkage methods, facilitating more informed decisions. Last but not least, the appropriate clustering model that was chosen based on the simulated study had been applied on the Juru River Basin water quality real dataset with the aim to verify and demonstrate

the appropriate clustering model. Based on the real data results, it is concluded that the condition of the Juru river basin during the months and years is considered to be the most contaminated in C1, the second most polluted in C3, the third most polluted in C4, the fourth polluted in C5, and the least polluted during the periods in C2. This is because C1 has the highest concentration levels of COD, BOD, and the lowest level of the DO parameter, followed by C3, C4, and C5. Meanwhile, C2 is the least polluted, as it has the lowest TSS and COD levels and high DO levels. This indicates that the river has low decomposition of organic and non-organic elements chemically and a low content of suspended solids. The high level of DO shows that the river contains a high amount of dissolved oxygen.

Moreover, based on the analysis of real data, high and low values of these water parameters as analyzed in previous chapter have their own meaning or interpretation about the river itself. COD involves the decomposition of organic and non-organic elements chemically. High COD level in river indicates high river pollution which could be resulted from the discharge of extensive organic waste and industry sewage (Hashim et al., 2019). Next, high TSS levels shows that there is high content of suspended solids such as sediment, silt that remain suspended in the water column which can reduce light penetration into the river, affecting aquatic plants and disrupting the habitat for aquatic organisms (Abdul Maulud et al., 2021). In addition, low DO levels indicates that the content of dissolved oxygen is low in the river that could be resulted from the industrial and domestic sewage disposal. The lower DO value indicates that the water source is highly contaminated (Hashim et al., 2021). Ammoniacal nitrogen, $\text{NH}_3\text{-NL}$ parameter refers to the presence of ammonia (NH_3) and ammonium ions (NH_4^+) in the water. Elevated levels of $\text{NH}_3\text{-NL}$ can result from agricultural runoff, untreated sewage, or industrial discharges (Baharudin,

Kassim, Mohd Imran & Wahab, 2021). Last but not least, high BOD concentrations in river also indicates that the river had been polluted. This is possibly due to natural plant decay, farming, fertilizer, septic systems and livestock wherein can also lead to stressful conditions for aquatic life (Masthurah et al., 2021). Therefore, overall, based on all the concentration levels of the five parameters, it can be concluded that water quality Juru river basin is severely contaminated throughout the year 2008 until 2017.

5.2 Significance of Study

The findings from this study yield several noteworthy contributions. First, in terms of the development of robustness theory within the field of clustering, the findings from simulation study had offered insights and knowledge regarding which clustering algorithm is suitable to be used based on the number of variables, sample size, percentage of outliers, and the type of data distributions. For example, *Spearman-average* clustering algorithm was the most appropriate choice when dealing with normal and contaminated distributions while *Winsorized-average* algorithm was the most suitable when dealing with non-normal Gamma distribution. This study also validated the efficacy of correlation-based distance in agglomerative hierarchical clustering. By comparing and evaluating the performances of the clustering algorithms by using average CCC value, this research demonstrated the robustness and reliability of the correlation-based distances in agglomerative hierarchical clustering approach.

Next, the focus on the application of correlation-based distances in this study had helped to provide guidelines for future researchers for selecting appropriate correlation-based distances especially when there are outliers in data. Based on the

simulation results, it can be seen that *Spearman-average* was the most appropriate clustering algorithm when dealing with data that contained outliers as it often displayed high CCC values across various number of variables and sample sizes. Furthermore, by employing correlation-based distance metrics, this study had facilitated the formation of clusters that reflect meaningful relationships between variables. Based on the findings on real data application, the *Spearman-average* algorithm had helped to identify the level of Juru river basin contamination within some periods based on the patterns of clusters formed from the concentration levels of the five water quality parameters. For instances, it was concluded that the condition of the Juru river basin during the months and years is considered to be the most contaminated in C1 as it had the highest concentration levels of COD, BOD, and the lowest level of the DO parameter while the river was the least polluted in C2 as it has the lowest TSS and COD levels and high DO levels.

Moreover, the findings from the real data application had also offered valuable information on the level of river contamination to the community such as industry players and residents in Juru area, stakeholders and government entities, such as the Penang State Government and the Department of Environment Malaysia. This information had the potential to assist them in addressing the crisis of contamination and water quality in the Juru river basin. For example, the high concentration levels of COD, BOD and low DO in C1 indicated that the river was polluted resulted from the high discharge of extensive organic waste, industry sewage, natural plant decay, farming, fertilizer, septic systems and livestock (Masthurah et al., 2021). Therefore, in order to protect the river from further contaminated, Malaysian government can formulate better policies and stringent regulatory measures while the community can

take actions by controlling the discharge of domestic or industry sewage into the river.

5.3 Limitation and Constraints faced in Study

Every study has its own limitations and constraints that shape the scope and depth of its findings. Like all research endeavours, this study also encountered its share of limitations and constraints which necessitated careful consideration throughout the investigation. First constraint was during the selection of the correlation-based distance and also the linkage method for agglomerative hierarchical clustering to be used in this study. While correlation-based distances and linkage method were explored, determining the most suitable methods to be used required careful consideration. Different methods could yield varying clustering results, influencing the interpretability and applicability of the findings.

Second constraint was in terms of computational challenges. During simulation study, calculating different correlation-based distances and linkage methods for large and high dimensionality datasets can be very computational time consuming which take longer days and even weeks to be completed. This is because the calculation of different methods for large high dimensionality datasets can be computationally intensive, requiring substantial processing power, time and memory.

Last but not least is the limitations on data availability and quality for real data applications. One of the primary constraints encountered was the availability and quality of the water quality parameter data. This thesis relied on existing datasets collected by local environmental agencies, which varied in completeness and consistency. Some parameters may have had sparse data points or were measured infrequently, limiting the robustness of the clustering analysis.

5.4 Recommendation for Future Researchers

The table of simulated results summary presented in this study can be used as a guideline for the future researchers to gain information on which correlation-based distance measure and linkage method is appropriate to be paired or used in constructing clustering algorithms and employ them on the data based on the types of data distributions, number of variables and sample sizes. They can also apply the clustering algorithms to the real data according to their respective fields of expertise.

Nonetheless, there is always space for future improvements. It is recommended for future studies to explore another robust correlation coefficients to be used as the distance measure or other type of linkage methods in constructing the clustering. This can help to expand greater knowledge pertaining to the clustering algorithms. Furthermore, it is suggested for the future researches to employ dimensionality reduction techniques such as feature selection or feature extraction before computing correlation-based distances. This might help to reduce the computational burden of the clustering algorithm and make it more effective when there is an increase in number of variables or sample sizes while preserving essential information.

Finally, the study demonstrated the practical application of the correlation-based distances in agglomerative hierarchical clustering approach in real-world data, particularly in the field of environmental science. It showcased the effectiveness of this approach in analysing datasets concerning water quality parameters. Therefore, future researchers are encouraged to expand upon these findings to investigate further interdisciplinary connections among water quality, ecology, hydrology, public health, and other related fields.

REFERENCES

- Abdul-Karim, B., & Sharif, S. (2019). An investigation of Juru River basin water quality using principal component analysis. *AIP Conference Proceedings*, 2138(050001), 1–4. <https://doi.org/10.1063/1.5121106>
- Abdul Maulud, K. N., Fitri, A., Wan Mohtar, W. H. M., Wan Mohd Jaafar, W. S., Zuhairi, N. Z., & Kamarudin, M. K. A. (2021). A study of spatial and water quality index during dry and rainy seasons at Kelantan River Basin, Peninsular Malaysia. *Arabian Journal of Geosciences*, 14(85), 1–19. <https://doi.org/10.1007/s12517-020-06382-8>
- Abdul Zali, M., Retnam, A., & Juahir, H. (2011). Spatial characterization of water quality using principal component analysis approach at Juru River Basin , Malaysia. *World Applied Sciences Journal*, 14, 55–59.
- Adams, F. V., Arseneau-Robar, T. J. M., Bonnell, T. R., Stead, S. M., & Teichroeb, J. A. (2021). Temporal patterns in the social network of core units in Rwenzori Angolan colobus monkeys: Effects of food availability and interunit dispersal. In *Ecology and Evolution* (Vol. 11, Issue 7, pp. 3251–3263). <https://doi.org/10.1002/ece3.7274>
- Aggarwal, C. C. (2015). Data Mining: The Textbook - Mining Discrete Sequences. *Springer International Publishing*, 746. <https://doi.org/10.1007/978-3-319-14142-8>
- Ahad, N. A., Zakaria, N. A., Abdullah, S., Yahaya, S. S. S., & Yusof, N. (2018). Robust correlation procedure via sn estimator. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(1–10), 115–118.
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, 18(3), 91–93. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Akogul, S., & Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications*, 21(3). <https://doi.org/10.3390/mca21030034>
- Al-badaii, F., Shuhaimi-othman, M., & Gasim, M. B. (2013). Water Quality Assessment of the Semeniyh River ,. *Journal of Chemistry*, 3216, 112–122.
- Al-Jawher, W. A. M., & Shaban, S. A. (2022). Clustering of Medical Images Using Multiwavelet Transform and K-Means Algorithms. *Journal Port Science Research*, 5(1), 35–42. <https://doi.org/10.36371/port.2022.1.5>
- Alam, A., Muqem, M., & Ahmad, S. (2021). Comprehensive review on Clustering Techniques and its application on High Dimensional Data. *International Journal of Computer Science & Network Security*, 21(6), 237–244.
- Alameddine, I., Kenney, M. A., Gosnell, R. J., & Reckhow, K. H. (2010). Robust Multivariate Outlier Detection Methods for Environmental Data. *Journal of Environmental Engineering*, 136(11), 1299–1304. [https://doi.org/10.1061/\(asce\)ee.1943-7870.0000271](https://doi.org/10.1061/(asce)ee.1943-7870.0000271)

- Albalate, A., & Suendermann, D. (2009). A combination approach to cluster validation based on statistical quantiles. *2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, IJCBS*, 549–555. <https://doi.org/10.1109/IJCBS.2009.116>
- Albalate, A., Suendermann, D., & Minker, W. (2011). On cluster validation for detecting the number of clusters in a data set. *International Journal on Artificial Intelligence Tools*, 20(5), 941–953. <https://doi.org/10.1142/S0218213011000334>
- Alfaro, J. L., & Ortega, J. F. (2009). A comparison of robust alternatives to Hotelling's T² control chart. *Journal of Applied Statistics*, 36(12), 1385–1396. <https://doi.org/10.1080/02664760902810813>
- Alkenani, A., & Yu, K. (2013). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation*, 83(4), 692–720. <https://doi.org/10.1080/00949655.2011.632775>
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1
- Almeida, J. A. S., Barbosa, L. M. S., Pais, A. A. C. C., & Formosinho, S. J. (2007). Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems*, 87(2), 208–217. <https://doi.org/10.1016/j.chemolab.2007.01.005>
- Alsaqr, A. M. (2021). Remarks on the Use of Pearson's and Spearman's Correlation Coefficients in Assessing Relationships in Ophthalmic Data. *African Vision and Eye Health*, 80(1), 1–10. <https://doi.org/10.4102/AVEH.V80I1.612>
- Amorim, D. R. C., & Hennig, C. (2015). Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information Sciences*, 324, 126–145. <https://doi.org/10.1016/j.ins.2015.06.039>
- Ankur, Y., & Singh, A. S. (2014). Outlier Analysis Using Frequent Pattern Mining – A Review. *International Journal of Computer Science and Information Technologies*, 5(1), 47–50.
- Arguelles, M., Benavides, C., & Fernández, I. (2014). A new approach to the identification of regional clusters: Hierarchical clustering on principal components. *Applied Economics*, 46(21), 2511–2519. <https://doi.org/10.1080/00036846.2014.904491>
- Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics*, 39(5), 316–327. <https://doi.org/10.1111/opo.12636>
- Arora, S., & Chana, I. (2014). A survey of clustering techniques for big data analysis. *Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technology Summit*, 59–65.

<https://doi.org/10.1109/CONFLUENCE.2014.6949256>

- Asuero, A. G., Sayago, A., & Gonzalez, A. G. (2006). The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36(1), 41–59. <https://doi.org/10.1080/10408340500526766>
- Aubert, A. H., Tavenard, R., Emonet, R., De Lavenne, A., Malinowski, S., Guyet, T., Quiniou, R., Odobez, J. M., Merot, P., & Gascuel-Odoux, C. (2013). Clustering flood events from water quality time series using Latent Dirichlet Allocation model. *Water Resources Research*, 49(12), 8187–8199. <https://doi.org/10.1002/2013WR014086>
- Babichev, S., Taif, M. A., Lytvynenko, V., & Osypenko, V. (2017). Criterial analysis of gene expression sequences to create the objective clustering inductive technology. *2017 IEEE 37th International Conference on Electronics and Nanotechnology, ELNANO 2017 - Proceedings*, 244–248. <https://doi.org/10.1109/ELNANO.2017.7939756>
- Badr, H. S., Zaitchik, B. F., & Dezfuli, A. K. (2015). A tool for hierarchical climate regionalization. *Earth Science Informatics*, 8(4), 949–958. <https://doi.org/10.1007/s12145-015-0221-7>
- Baharudin, F., Kassim, J., Imran, S. N. M., & Wahab, M. A. (2021). Water Quality Index (WQI) classification of rivers in agriculture and aquaculture catchments. *IOP Conference Series: Earth and Environmental Science*, 646(012023). <https://doi.org/10.1088/1755-1315/646/1/012023>
- Baker, F. B. (1974). Stability of two hierarchical grouping techniques case I: Sensitivity to data errors. *Journal of the American Statistical Association*, 69(346), 440–445. <https://doi.org/10.1080/01621459.1974.10482971>
- Beckman, R. J., & Cook, R. D. (1983). Outlier s. *Technometrics*, 25(2), 119–149. <https://doi.org/10.1080/00401706.1983.10487840>
- Bhaumik, D. K., Kapur, K., & Gibbons, R. D. (2009). Testing Parameters of a Gamma Distribution for Small Samples. *Technometrics*, 51(3), 326–334. <https://doi.org/10.1198/tech.2009.07038>
- Bindra, K., Mishra, A., & Suryakant. (2019). Effective data clustering algorithms. In *Advances in Intelligent Systems and Computing* (Vol. 742). Springer Singapore. https://doi.org/10.1007/978-981-13-0589-4_39
- Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 83(3), 377–388. <https://doi.org/10.1037/0033-2909.83.3.377>
- Blashfield, R. K., & Aldenderfer, M. S. (1988). The Methods and Problems of Cluster Analysis. *Handbook of Multivariate Experimental Psychology*, 1, 447–473. https://doi.org/10.1007/978-1-4613-0893-5_14
- Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S. F., Dunning, M. J., Barbosa-Morais, N. L., Teschendorff, A. E., Green, A. R., Ellis, I. O., Tavaré,

- S., Caldas, C., & Miska, E. A. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology*, 8(10). <https://doi.org/10.1186/gb-2007-8-10-r214>
- Boyko, N. I., & Tkachyk, O. A. (2023). Hierarchical clustering algorithm for dendrogram construction and cluster counting. *Informatics and Mathematical Methods in Simulation*, 13(1–2), 5–15. <https://doi.org/10.15276/imms.v13.no1-2.5>
- Bratchell, N. (2003). Cluster Analysis. *Swiss Journal of Psychology*, 62(3), 202–202. <https://doi.org/10.1024//1421-0185.62.3.202h>
- Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations)? *Journal of Computational and Graphical Statistics*, 7(4), 434–455. <https://doi.org/10.1080/10618600.1998.10474787>
- Bu, J., Liu, W., Pan, Z., & Ling, K. (2020). Comparative study of hydrochemical classification based on different hierarchical cluster analysis methods. *International Journal of Environmental Research and Public Health*, 17(9515), 1–23. <https://doi.org/10.3390/ijerph17249515>
- Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, 3(1), 1–27. <https://doi.org/10.1080/03610927408827101>
- Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2015). Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data*, 10(1), 1–51. <https://doi.org/10.1145/2733381>
- Carvalho, P. R., Munita, C. S., & Lapolli, A. L. (2019). Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient. *Brazilian Journal of Radiation Sciences*, 7(2A), 1–14. <https://doi.org/10.15392/bjrs.v7i2a.668>
- Chaakraborty, P., Sen, S., & Maity, S. (2020). A Comparative Study of Clustering Models in Data Mining. *A Multidisciplinary Journal*, 1(1), 28–35.
- Chauhan, T., Rawat, S., Malik, S., & Singh, P. (2021). Supervised and Unsupervised Machine Learning based Review on Diabetes Care. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 581–585. <https://doi.org/10.1109/ICACCS51430.2021.9442021>
- Chen, J., Ng, Y. K., Lin, L., Zhang, X., & Li, S. (2023). On triangle inequalities of correlation - based distances for gene expression profiles. *BMC Bioinformatics*, 24(40), 1–16. <https://doi.org/10.1186/s12859-023-05161-y>
- Chen, P., Buis, K., & Zhao, X. (2023). A comprehensive toolbox for the gamma distribution: The gammadist package. *Journal of Quality Technology*, 55(1), 57–69. <https://doi.org/10.1080/00224065.2022.2053794>
- Chen, Z., Yang, W., Liu, Q., Yang, J. Y., Li, J., & Yang, M. Q. (2014). A new

- statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics*, 15(17), 1–7. <https://doi.org/10.1186/1471-2105-15-S17-S3>
- Chiang, M. M.-T., & Mirkin, B. (2010). Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads. *Journal of Classification*, 27, 3–40. <https://doi.org/10.1007/s00357-010->
- Chok, N. S. (2008). Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data. In *Thesis*.
- Choudhury, J., & Ashraf, F. Bin. (2022). An Analysis of Different Distance-Linkage Methods for Clustering Gene Expression Data and Observing Pleiotropy: Empirical Study. *JMIR Bioinformatics and Biotechnology*, 3(1), e30890. <https://doi.org/10.2196/30890>
- Colwell, D. J., & Gillet, J. R. (1982). Spearman versus Kendall. *Mathematical Association*, 66(438), 307–309.
- Componation, P. J., & Byrd, J. (2000). Utilizing cluster analysis to structure concurrent engineering teams. *IEEE Transactions on Engineering Management*, 47(2), 269–280. <https://doi.org/10.1109/17.846793>
- Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, 19, 497–515. <https://doi.org/10.1007/s10260-010-0142-z>
- D'Urso, P., & Vitale, V. (2020). A robust hierarchical clustering for georeferenced data. *Spatial Statistics*, 35, 100407. <https://doi.org/10.1016/j.spasta.2020.100407>
- Dabhi, D. P., & Patel, M. R. (2016). Extensive Survey on Hierarchical Clustering Methods in Data Mining. *International Research Journal of Engineering and Technology (IRJET)*, 03(11), 659–665. <https://www.irjet.net/archives/V3/i11/IRJET-V3I11115.pdf>
- Dahal, S. (2015). *Effect of Different Distance Measures in Result of Cluster Analysis*.
- Dange, B. J., Kumar Mishra, P., Metre, K. V, Gore, S., Laxnamrao Kurkute, S., Khodke, H. E., & Gore, S. (2023). Grape Vision: A CNN-Based System for Yield Component Analysis of Grape Clusters. *International Journal of Intelligent Systems and Applications in Engineering IJISAE*, 11(9s), 239–244. <https://orcid.org/0000-0003-1814-591>,
- Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*(2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>
- Davis, R. A., Matsui, M., Mikosch, T., & Wan, P. (2016). Applications of distance correlation to time series. *Bernoulli*, 24(4A), 2429–2460. <https://doi.org/10.3150/17-BEJ955>

- de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermit, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: A comparative study. *BMC Bioinformatics*, 9, 1–14. <https://doi.org/10.1186/1471-2105-9-497>
- Derya, K. O., & Canan, H. (2012). Control charts for skewed distributions: Weibull, gamma, and lognormal. *Metodoloski Zvezki*, 9(2), 95–106.
- Di Cesare, F., Luchinat, C., Tenori, L., & Saccenti, E. (2022). Age- and Sex-Dependent Changes of Free Circulating Blood Metabolite and Lipid Abundances, Correlations, and Ratios. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences*, 77(5), 918–926. <https://doi.org/10.1093/gerona/ghab335>
- Ding, C., Zhou, D., He, X., & Zha, H. (2006). R1-PCA: Rotational invariant L1-norm principal component analysis for robust subspace factorization. *ACM International Conference Proceeding Series*, 148, 281–288. <https://doi.org/10.1145/1143844.1143880>
- Dogan, A., & Birant, D. (2022). K-centroid link: a novel hierarchical clustering linkage method. *Applied Intelligence*, 52(5), 1–24. <https://doi.org/10.1007/s10489-021-02624-8>
- Dooley, K. (2002). Simulation Research Methods. In J. Baum (Ed.), *Companion to Organizations* (pp. 829–848). Blackwell.
- Doudney, C. O., & Rinaldi, C. N. (1984). Modification of UV-induced mutation frequency and cell survival of Escherichia coli B/r WP2 trp E65 by treatment before irradiation. *Journal of Bacteriology*, 160(1), 233–238. <https://doi.org/10.1128/jb.160.1.233-238.1984>
- Downs, G. M., & Barnard, J. M. (2002). Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry* (Vol. 18). <https://doi.org/10.1002/0471433519.ch1>
- Du, X., Shao, F., Wu, S., Zhang, H., & Xu, S. (2017). Water quality assessment with hierarchical cluster analysis based on Mahalanobis distance. *Environmental Monitoring and Assessment*, 189(7). <https://doi.org/10.1007/s10661-017-6035-y>
- Dubes, R. C. (1987). How many clusters are best? - An experiment. *Pattern Recognition*, 20(6), 645–663. [https://doi.org/10.1016/0031-3203\(87\)90034-3](https://doi.org/10.1016/0031-3203(87)90034-3)
- Dubreuil, P., & Charcosset, A. (1998). Genetic diversity within and among maize populations: A comparison between isozyme and nuclear RFLP loci. *Theoretical and Applied Genetics*, 96, 577–587. <https://doi.org/10.1007/s001220050776>
- Dunn, J. C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4(1), 95–104.
- Ebrahimi, P., Basirat, M., Yousefi, A., Nekmahmud, M., Gholampour, A., & Fekete-farkas, M. (2022). Social Networks Marketing and Consumer Purchase Behavior: The Combination of SEM and Unsupervised Machine Learning

- Approaches. *Big Data and Cognitive Computing*, 6(2), 1–18. <https://doi.org/10.3390/bdcc6020035>
- El-Hashash, E. F., & Shiekh, R. H. A. (2022). A Comparison of the Pearson, Spearman Rank and Kendall Tau Correlation Coefficients Using Quantitative Variables. *Asian Journal of Probability and Statistics*, 20(3), 36–48. <https://doi.org/10.9734/ajpas/2022/v20i3425>
- Emmendorfer, L. R., & de Paula Canuto, A. M. (2021). A generalized average linkage criterion for Hierarchical Agglomerative Clustering. *Applied Soft Computing*, 100. <https://doi.org/10.1016/J.ASOC.2020.106990>
- Erisoglu, M., & Sakallioğlu, S. (2010). An Investigation of Effects on Hierarchical Clustering of Distance Measurements. *Selcuk Journal of Applied Mathematics, Special Issue*, 39–53.
- Ertoz, L., Steinbach, M., & Kumar, V. (2003). *Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data*. May, 47–58. <https://doi.org/10.1137/1.9781611972733.5>
- Estivill-Castro, V. (2002). Why so many clustering algorithms. *ACM SIGKDD Explorations Newsletter*, 4(1), 65–75. <https://doi.org/10.1145/568574.568575>
- Eszergar-Kiss, D., & Caesar, B. (2017). Definition of user groups applying Ward's method. In *Transportation Research Procedia* (Vol. 22, pp. 25–34). <https://doi.org/10.1016/j.trpro.2017.03.004>
- Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis. In *Wiley Series in Probability and Statistics* (5th Editio). John Wiley & Sons, Ltd. <https://doi.org/10.1007/BF00154794>
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing*, 1–12. <https://doi.org/10.1109/TETC.2014.2330519>
- Faroh, W. N., Tajuddin, R., Narimawati, U., Affandi, A., & Priadana, S. (2022). Clustering Hierarchical and Non-Hierarchical Algorithm : Clustering of 21st Century Hr Skillsi In SMKN Students In Indonesia. *International Journal of Education, Information Technology and Others (IJEIT)*, 5(2), 158–165. <https://doi.org/10.5281/zenodo.6418031>
- Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3), 279–285. <https://doi.org/10.2307/2412324>
- Fernandes, A. A. R. (2021). Comparison of Cluster and Linkage Validity Indices in Integrated Cluster Analysis with Structural Equation Modeling War-PLS Approach. *Journal of Hunan University*, 48(4), 88–99. <http://www.jonuns.com/index.php/journal/article/view/542/539>
- Ferreira, L., & Hitchcock, D. B. (2009). A comparison of hierarchical methods for clustering functional data. *Paper Knowledge*, 1–26.

- Filzmoser, P., Serneels, S., Maronna, R., & Van Espen, P. J. (2009). Robust Multivariate Methods in Chemometrics. In *Comprehensive Chemometrics* (Vol. 3, pp. 681–722). <https://doi.org/10.1016/B978-044452701-1.00113-7>
- Flores, J. L., Inza, I., Larrañaga, P., & Calvo, B. (2013). A new measure for gene expression biclustering based on non-parametric correlation. *Computer Methods and Programs in Biomedicine*, 112, 367–397. <https://doi.org/10.1016/j.cmpb.2013.07.025>
- Fovell, R. G. (1997). Consensus clustering of U.S. temperature and precipitation data. *Journal of Climate*, 10(6), 1405–1427. [https://doi.org/10.1175/1520-0442\(1997\)010<1405:CCOUST>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<1405:CCOUST>2.0.CO;2)
- Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383), 553–569.
- Ge, J., & Tibshirani, R. (2022). *Weakest link pruning of a dendrogram* (pp. 1–18). <https://arxiv.org/abs/2212.05367v1>
- Gibbons, F. D., & Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. In *Genome Research* (Vol. 12, Issue 10, pp. 1574–1581). <https://doi.org/10.1101/gr.397002>
- Gogtay, N. J., & Thatte, U. M. (2017). Principles of correlation analysis. *Journal of Association of Physicians of India*, 65(MARCH), 78–81.
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *Journal of Experimental Education*, 74(3), 249–266. <https://doi.org/10.3200/JEXE.74.3.249-266>
- Greenlaw, R., & Kantabutra, S. (2014). Survey of Clustering. *International Journal of Information Retrieval Research*, 3(2), 1–29. <https://doi.org/10.4018/ijirr.2013040101>
- Guha, S., Rastogi, R., & Shim, K. (1998). CURE: An efficient clustering algorithm for large databases. *SIGMOD Record*, 27(2), 73–84. <https://doi.org/10.1145/276305.276312>
- Gulagiz, F. K., & Suhap, S. (2017). Comparison of Hierarchical and Non-Hierarchical Clustering Algorithms. *International Journal of Computer Engineering and Information Technology*, 9(1), 6–14. www.ijceit.org
- Haddad, F. S., Syed-Yahaya, S. S., & Alfaro, J. L. (2012). Alternative Hotelling's T² charts using winsorized modified one-step M-estimator. *Quality and Reliability Engineering International*, 1–11. <https://doi.org/10.1002/qre.1407>
- Hadi, A. S. (2022). A new distance between multivariate clusters of varying locations, elliptical shapes, and directions. *Pattern Recognition*, 129, 108780. <https://doi.org/10.1016/j.patcog.2022.108780>
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). Clustering algorithms and

- validity measures. *Proceedings of the International Conference on Scientific and Statistical Database Management, SSDBM*, 3–22. <https://doi.org/10.1109/ssdm.2001.938534>
- Halkidi, Maria, Batistakis, Y., & Vazirgiannis, M. (2002). Cluster Validity Methods: Part I. *SIGMOD Record*, 31(2), 40–45.
- Hamzah, N. A., Kek, S. L., & Saharan, S. (2017). The Performance of K-Means and K-Modes Clustering to Identify Cluster in Numerical Data. *Journal of Science and Technology*, 9(3), 25–32.
- Hands, S., & Everitt, B. (1987). A Monte Carlo Study of the Recovery of Cluster Structure in Binary Data by Hierarchical Clustering Techniques. *Multivariate Behavioral Research*, 22(2), 235–243. https://doi.org/10.1207/s15327906mbr2202_6
- Hartigan, J. A. (1973). Clustering. *Annual Review of Biophysics and Bioengineering*, 2, 81–101.
- Hashim, M., Nor, S. S. M., Nayan, N., Mahat, H., Saleh, Y., See, K. L., & Norkhaidi, S. B. (2019). Analysis of Well Water Quality in the District of Pasir Puteh, Kelantan, Malaysia. *IOP Conference Series: Earth and Environmental Science*, 286(012021), 1–10. <https://doi.org/10.1088/1755-1315/286/1/012021>
- Hashim, Mohmadisa, Nayan, N., Setyowati, D. L., Said, Z. M., Mahat, H., & Saleh, Y. (2021). Analysis of Water Quality Trends Using the Mann-Kendall Test and Sen's Estimator of Slope in a Tropical River Basin. *Pollution*, 7(4), 933–942. <https://doi.org/10.22059/POLL.2021.325794.1118>
- Hastings, C., Mosteller, F., Tukey, J. W., & Winsor, C. P. (1947). Low Moments for Small Samples: A Comparative Study of Order Statistics. *The Annals of Mathematical Statistics*, 18(3), 413–426. <https://doi.org/10.1214/aoms/1177730388>
- Hawkins, D. M. (1980). Identification of Outliers. In *Identification of Outliers*. <https://doi.org/10.1007/978-94-015-3994-4>
- Heuvel, E. van den, & Zhan, Z. (2022). Myths About Linear and Monotonic Associations: Pearson's r , Spearman's ρ , and Kendall's τ . *The American Statistician*, 76(1), 44–52. <https://doi.org/10.1080/00031305.2021.2004922>
- Ho, J. Y., Afan, H. A., El-Shafie, A. H., Koting, S. B., Mohd, N. S., Jaafar, W. Z. B., Lai Sai, H., Malek, M. A., Ahmed, A. N., Mohtar, W. H. M. W., Elshorbagy, A., & El-Shafie, A. (2019). Towards a time and cost effective approach to water quality index class prediction. *Journal of Hydrology*, 1–46. <https://doi.org/10.1016/j.jhydrol.2019.05.016>
- Hu, W., & Pan, Q. he. (2015). Data clustering and analyzing techniques using hierarchical clustering method. *Multimedia Tools and Applications*, 74(19), 8495–8504. <https://doi.org/10.1007/s11042-013-1611-9>
- Irani, J., Pise, N., & Phatak, M. (2016). Clustering Techniques and the Similarity

- Measures used in Clustering: A Survey. *International Journal of Computer Applications*, 134(7), 9–14. <https://doi.org/10.5120/ijca2016907841>
- Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., & van Diepen, M. (2021). Conducting correlation analysis: important limitations and pitfalls. *Clinical Kidney Journal*, 14(10), 2332–2337. <https://doi.org/10.1093/ckj/sfab085>
- Jardim, S., & Mora, C. (2022). Customer reviews sentiment-based analysis and clustering for market-oriented tourism services and products development or positioning. *Procedia Computer Science*, 196(2022), 199–206. <https://doi.org/10.1016/j.procs.2021.12.006>
- Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. In *Research gate*. <https://doi.org/10.13140/RG.2.2.11388.90240>
- Jaskowiak, P. A., Campello, R. J., & Costa, I. G. (2014). On the Selection of Appropriate Proximity Measurement for Gene Expression Data. *BMC Bioinformatics*, 15. <https://doi.org/10.11648/j.ijbmr.20170505.11>
- Johnson, R.A. and Wichern, D.W. (2015) *Applied Multivariate Statistical Analysis*, 6th Edition, Pearson, India.
- K.Mohbey, K., & S. Thakur, G. (2013). An Experimental Survey on Single Linkage Clustering. *International Journal of Computer Applications*, 76(17), 6–11. <https://doi.org/10.5120/13337-0327>
- Kangane, P., Joshi, V., Kacker, A., & Jain, M. (2021). A Comprehensive Survey of Various Clustering Paradigms. *International Journal of Engineering Applied Sciences and Technology*, 5(11), 255–261. <https://doi.org/10.33564/ijeast.2021.v05i11.042>
- Karim, B. A., Sharif, S., Zakaria, H., & Azis, T. M. F. (2019). Influential water quality parameter in Juru River basin using PCA. *Test Engineering and Management*, 81, 819–825.
- Kassambara, A. (2015). Multivariate Analysis 1: Practical Guide To Cluster Analysis in R (Unsupervised Machine Learning). In *STHDA*.
- Kaufman, L., & Rousseeuw, P. J. (1990). Agglomerative Nesting (Program AGNES). In *Finding Groups in Data: An Introduction to Cluster Analysis* (pp. 199–252). <https://doi.org/10.1002/9780470316801.ch5>
- Kazi, A., & Kurian, P. D. T. (2014). A Survey of Data Clustering Techniques. *International Journal of Engineering Research & Technology*, 3(10), 174–177.
- Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1–2), 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>
- Khoo, M. B. C., Saha, S., Teh, S. Y., Haq, A., & Lee, H. C. (2020). The median control chart for process monitoring in short production runs. *Communications*

in *Statistics: Simulation and Computation*, 1–16.
<https://doi.org/10.1080/03610918.2020.1783557>

- Kim, H., Robert, C. P., & Casella, G. (2000). Monte Carlo Statistical Methods. *Technometrics*, 42(4), 430. <https://doi.org/10.2307/1270959>
- Kim, Y., Kim, T. H., & Ergün, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters*, 13, 243–257. <https://doi.org/10.1016/j.frl.2014.12.005>
- Kimes, P. K., Liu, Y., Neil Hayes, D., & Marron, J. S. (2017). Statistical significance for hierarchical clustering. *Biometrics*, 73(3), 811–821. <https://doi.org/10.1111/biom.12647>
- Kothari, R., & Pitts, D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20(4), 405–416. [https://doi.org/10.1016/S0167-8655\(99\)00008-2](https://doi.org/10.1016/S0167-8655(99)00008-2)
- Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wilcox, R. R. (2006). Multiple comparison procedures, trimmed means and transformed statistics. *Journal of Modern Applied Statistical Methods*, 5(1), 43–65. <https://doi.org/10.22237/jmasm/1146456300>
- Kowalczyk, M., & Nizankowski, C. (2017). Comparative Analysis of Machinability of Nitinol Alloy Using Weighted Radar Diagram. *Management and Production Engineering Review*, 8(4), 74–81. <https://doi.org/10.1515/mper-2017-0038>
- Kozak, M., Krzanowski, W., & Tartanus, M. (2012). Use of the correlation coefficient in agricultural sciences: Problems, pitfalls and how to deal with them. *Anais Da Academia Brasileira de Ciencias*, 84(4), 1147–1156. <https://doi.org/10.1590/S0001-37652012000400029>
- Kuiper, F. K., & Fisher, L. (2011). A Monte Carlo Comparison of Six Clustering Procedures. *Biometrics*, 31(3), 777–783.
- Kulesz, P. A., Tian, S., Juranek, J., Fletcher, J. M., & Francis, D. J. (2015). Relations between volumetric measures of brain structure and attentional function in Spina bifida: Utilization of robust statistical approaches. *Neuropsychology*, 29(2), 212–225. <https://doi.org/10.1037/neu0000166>
- Kulkarni, A., Tokekar, V., & Kulkarni, P. (2015). Discovering context of labeled text documents using context similarity coefficient. *Procedia Computer Science*, 49(1), 118–127. <https://doi.org/10.1016/j.procs.2015.04.235>
- Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data*, 3(13), 1–11. <https://doi.org/10.1186/s40537-016-0046-3>
- Kumar, V., Chhabra, J. K., & Kumar, D. (2014). Impact of distance measures on the performance of clustering algorithms. *Advances in Intelligent Systems and Computing*, 243, 183–190. https://doi.org/10.1007/978-81-322-1665-0_17

- Kurtzman, R. (2008). *The Advantages of a Biweight Metric in Clustering Microarray Data*.
- Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. In *Korean Journal of Anesthesiology* (Vol. 70, Issue 4, pp. 407–411). <https://doi.org/10.4097/kjae.2017.70.4.407>
- Lalitha, Y. S. (2014). An Advanced Agglomerative Hierarchical Clustering Methods on Aquatic Data Set. *International Journal of Aquatic Science*, 05(02), 215–223.
- Lee, S., Kim, J., Hwang, J., Lee, E. J., Lee, K. J., Oh, J., Park, J., & Heo, T. Y. (2020). Clustering of time series water quality data using dynamic time warping: A case study from the Bukhan River water quality monitoring network. *Water (Switzerland)*, 12(9). <https://doi.org/10.3390/w12092411>
- Li, Q., & Kim, B. M. (2003). Clustering approach for hybrid recommender system. *Proceedings - IEEE/WIC International Conference on Web Intelligence, WI 2003*, 33–38. <https://doi.org/10.1109/WI.2003.1241167>
- Liu, H., Li, J., Wu, Y., & Fu, Y. (2019). Clustering With Outlier Removal. *IEEE Transactions on Knowledge and Data Engineering*, 33(6), 2369–2379. <https://doi.org/10.1109/tkde.2019.2954317>
- Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of internal clustering validation measures. *IEEE International Conference on Data Mining, ICDM*, 911–916. <https://doi.org/10.1109/ICDM.2010.35>
- Lizbetinova, L., Starchon, P., Lorincova, S., Weberova, D., & Prusa, P. (2019). Application of cluster analysis in marketing communications in small and medium-sized enterprises: An empirical study in the Slovak Republic. *Sustainability (Switzerland)*, 11(8). <https://doi.org/10.3390/su11082302>
- Lopes, H. E. G., & Gosling, M. de S. (2021). Cluster Analysis in Practice: Dealing with outliers in Managerial Research. *Journal of Contemporary Administration*, 25(1), 1–19.
- Lyra, G. B., Oliveira-Júnior, J. F., & Zeri, M. (2014). Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Alagoas state, Northeast of Brazil. *International Journal of Climatology*, 34(13), 3546–3558. <https://doi.org/10.1002/joc.3926>
- Ma, X., & Dhavala, S. (2018). *Hierarchical Clustering with Prior Knowledge* (pp. 1–9). <http://arxiv.org/abs/1806.03432>
- Mahesh Batta. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJSR)*, October. <https://doi.org/10.21275/ART20203995>
- Mair, P., & Wilcox, R. R. (2019). *Robust statistical methods in R using the WRS2 package*.

- Mamun, A. Al, Aseltine, R., & Rajasekaran, S. (2016). Efficient record linkage algorithms using complete linkage clustering. *PLoS ONE*, 11(4), 1–21. <https://doi.org/10.1371/journal.pone.0154446>
- Mangiameli, P., Chen, S. K., & West, D. (1996). A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research*, 93, 402–417. [https://doi.org/10.1016/0377-2217\(96\)00038-0](https://doi.org/10.1016/0377-2217(96)00038-0)
- Manoharan, D. S. (2020). Performance analysis of clustering based image segmentation and optimization methods. *Journal of Innovative Image Processing (JIIP)*, 02(01), 14–24. <https://doi.org/https://doi.org/10.36548/jiip.2020.1.002>
- Masthurah, A., Juahir, H., & Mohd Zanuri, N. B. (2021). Case study malaysia: Spatial water quality assessment of Juru, Kuantan and Johor river basins using environmetric techniques. *Journal of Survey in Fisheries Sciences*, 7(2), 19–40. <https://doi.org/10.18331/SFS2021.7.2.2>
- Meena, K. S., & Suriya, S. (2020). A Survey on Supervised and Unsupervised Learning Techniques. In *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications* (pp. 909–923). https://doi.org/10.1007/978-3-030-24051-6_85
- Mei-Wo, Y., & Mahmood, Z. U. W. (2020). Estimation of sediment accumulation rate at different season in Juru River and Perai River (Perai Industrial Area, Penang, Malaysia). *Jurnal Sains Nuklear Malaysia*, 32(2), 8–22.
- Mei, F. F., Chuan, O. M., & Ling, L. C. (2016). Determination of Total Organic Carbon Concentration in Surficial Sediments of Sungai Pinang, Penang, Malaysia. *Middle-East Journal of Scientific Research*, 24(4), 1470–1476. <https://doi.org/10.17576/mjas-2016-2006-10>
- Mercioni, M. A., & Holban, S. (2018). Evaluating hierarchical and non-hierarchical grouping for develop a smart system. *2018 13th International Symposium on Electronics and Telecommunications, ISETC 2018 - Conference Proceedings*, 1–4. <https://doi.org/10.1109/ISETC.2018.8583997>
- Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3), 325–342. <https://doi.org/10.1007/BF02293907>
- Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2), 159–179.
- Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, 5(2), 181–204. <https://doi.org/10.1007/BF01897163>
- Mimmack, G. M., Mason, S. J., & Galpin, J. S. (2001). Choice of distance matrices in cluster analysis: Defining regions. *Journal of Climate*, 14(12), 2790–2797. [https://doi.org/10.1175/1520-0442\(2001\)014<2790:CODMIC>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<2790:CODMIC>2.0.CO;2)

- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38, 2074–2102. <https://doi.org/10.1002/sim.8086>
- Mukaka, M. M. (2012). Statistic corner: A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal*, 24(3), 69–71.
- Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1), 86–97. <https://doi.org/10.1002/widm.53>
- Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *Journal of Classification*, 31(3), 274–295. <https://doi.org/10.1007/s00357-014-9161-z>
- Nagi, S., Bhattacharyya, D. K., & Kalita, J. K. (2011). Gene expression data clustering analysis: A survey. *Proceedings - 2011 2nd National Conference on Emerging Trends and Applications in Computer Science, NCETACS-2011*, 1–12. <https://doi.org/10.1109/NCETACS.2011.5751377>
- Naubi, I., Zardari, N. H., Shirazi, S. M., Ibrahim, N. F. B., & Baloo, L. (2016). Effectiveness of water quality index for monitoring Malaysian river water quality. *Polish Journal of Environmental Studies*, 25(1), 231–239. <https://doi.org/10.15244/pjoes/60109>
- Newson, R. (2002). Parameters behind “Nonparametric” Statistics: Kendall's tau, Somers' D and Median Differences. *The Stata Journal*, 2(1), 45–64. <https://doi.org/10.1177/1536867x0200200103>
- Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review*, 52(1), 77–124. <https://doi.org/10.1007/s10462-018-09679-z>
- Nies, H. W., Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., Sinnott, R. O., Napis, S., Chamoso, P., Omatu, S., & Corchado, J. M. (2019). A review of computational methods for clustering genes with similar biological functions. *Processes*, 7(9), 1–18. <https://doi.org/10.3390/pr7090550>
- Nisha, & Kaur, P. J. (2016). Cluster quality based performance evaluation of hierarchical clustering method. *Proceedings on 2015 1st International Conference on Next Generation Computing Technologies, NGCT 2015, September*, 649–653. <https://doi.org/10.1109/NGCT.2015.7375201>
- Nowak-Brzezinska, A., & Gaibei, I. (2022). How the Outliers Influence the Quality of Clustering? *Entropy*, 24(917), 1–16. <https://doi.org/10.3390/e24070917>
- Ochuko, T. K., Abdullah, S., Zain, Z., & Yahaya, S. S. S. (2016). The Power of the Test for the Winsorized Modified Alexander-Govern Test. *ComTech: Computer, Mathematics and Engineering Applications*, 7(4), 307. <https://doi.org/10.21512/comtech.v7i4.3764>

- Oyelade, J., Isewon, I., Oladipupo, O., Emebo, O., Omogbadegun, Z., Aromolaran, O., Uwoghien, E., Olaniyan, D., & Olawole, O. (2019). Data Clustering: Algorithms and Its Applications. *Proceedings - 2019 19th International Conference on Computational Science and Its Applications, ICCSA 2019*, ii, 71–81. <https://doi.org/10.1109/ICCSA.2019.000-1>
- Palacio-Nino, J.-O., & Berzal, F. (2019). *Evaluation Metrics for Unsupervised Learning Algorithms*. <http://arxiv.org/abs/1905.05667>
- Panda, U. C., Sundaray, S. K., Rath, P., Nayak, B. B., & Bhatta, D. (2006). Application of factor and cluster analysis for characterization of river and estuarine water systems - A case study: Mahanadi River (India). *Journal of Hydrology*, 331(3–4), 434–445. <https://doi.org/10.1016/j.jhydrol.2006.05.029>
- Pandove, D., Goel, S., & Rani, R. (2018). Systematic review of clustering high-Dimensional and large datasets. *ACM Transactions on Knowledge Discovery from Data*, 12(2). <https://doi.org/10.1145/3132088>
- Panji, E., Gruenwald, L., Leal, E., Nguyen, C., & Silvia, S. (2022). A survey on outlier explanations. In *VLDB Journal* (pp. 977–1008). <https://doi.org/10.1007/s00778-021-00721-1>
- Patibandla, R. S. M. L., & Veeranjanyulu, N. (2018). Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria. *Arabian Journal for Science and Engineering*, 43(8), 4379–4390. <https://doi.org/10.1007/s13369-017-3036-7>
- Patrik, D. (2005). *How does gene expression clustering work ?* 23(12), 1499–1501.
- Pearson, K. (1896). VII. *Mathematical Contributions to the Theory of Evolution - III. Regression, Heredity and Panmixia*.
- Pei, J., Zhang, X., Cho, M., Wang, H., & Yu, P. S. (2003). MaPle: A fast algorithm for maximal pattern-based clustering. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 1(c), 259–266. <https://doi.org/10.1109/icdm.2003.1250928>
- Prasetyo, H., & Purwianti, A. (2015). Comparison of Distance and Dissimilarity Measures for Clustering Data with Mix Attribute Types. *International Conference on Information Technology, Computer, and Electrical Engineering: Green Technology and Its Applications for a Better Future, ICITACEE*, 276–280. <https://doi.org/10.1109/ICITACEE.2014.7065756>
- Puth, M. T., Neuhauser, M., & Ruxton, G. D. (2014). Effective use of Pearson's product-moment correlation coefficient. *Animal Behaviour*, 93, 183–189. <https://doi.org/10.1016/j.anbehav.2014.05.003>
- Puth, M. T., Neuhauser, M., & Ruxton, G. D. (2015). Effective use of Spearman's and Kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102, 77–84. <https://doi.org/10.1016/j.anbehav.2015.01.010>

- Qian, Y., Zhang, G., & Zhang, K. (2004). FACADE : A Fast and Effective Approach to the Discovery. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 2–3.
- Rai, P., & Singh, S. (2010). A Survey of Clustering Techniques. *International Journal of Computer Applications*, 7(12), 1–5. <https://doi.org/10.5120/1326-1808>
- Rajalingam, N., & Ranjini, K. (2011). Hierarchical Clustering Algorithm- A Comparative Study. *International ...*, 19(3), 42–46. <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Hierarchical+Clustering+Algorithm+-+A+Comparative+Study#0%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Hierarchical+Clustering+Algorithm+-+A+Comparative+Study#0>
- Rajesh, G., & Chaturvedi, A. (2019). Correlation Analysis of Multimodal Sensor Data in Environmental Sensor Networks. *International Symposium on Advanced Networks and Telecommunication Systems, ANTS*, 3–8. <https://doi.org/10.1109/ANTS47819.2019.9118157>
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336), 846–850. <https://doi.org/10.1080/01621459.1971.10482356>
- Rani, Y., & Rohil, H. (2013). A study of hierarchical clustering algorithms. *International Journal of Information and Computation Technology.*, 3(10), 1115–1122.
- Raveh, A. (1986). On Measures of Monotone Association. *The American Statistician*, 40(2), 117–123.
- Rehm, F., Klawonn, F., & Kruse, R. (2007). A novel approach to noise clustering for outlier detection. *Soft Computing*, 11(5), 489–494. <https://doi.org/10.1007/s00500-006-0112-4>
- Rincon, F., Johnson, B., Crossa, J., & Taba, S. (1996). Cluster analysis, an approach to sampling variability in maize accessions. *Maydica*, 41, 307–316.
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. In *PLoS ONE* (Vol. 14, Issue 1). <https://doi.org/10.1371/journal.pone.0210236>
- Rohlf, J. (1970). Adaptive Hierarchical Clustering Schemes. *Systematic Zoology*, 19(1), 58–82.
- Ros, F., & Guillaume, S. (2019). A hierarchical clustering algorithm and an improvement of the single linkage criterion to deal with noise. *Expert Systems with Applications*, 128, 96–108. <https://doi.org/10.1016/j.eswa.2019.03.031>
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and

- validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 73–79. <https://doi.org/10.1002/widm.2>
- Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. *Journal of Classification*, 35(2), 345–366. <https://doi.org/10.1007/s00357-018-9259-9>
- Saad, F. H. (2012). Comparison of Hierarchical Agglomerative Algorithms for Clustering Medical Documents. *International Journal of Software Engineering & Applications*, 3(3), 1–15. <https://doi.org/10.5121/ijsea.2012.3301>
- Sajana, T., Sheela Rani, C. M., & Narayana, K. V. (2016). A survey on clustering techniques for big data mining. *Indian Journal of Science and Technology*, 9(3), 1–12. <https://doi.org/10.17485/ijst/2016/v9i3/75971>
- Saracli, S., & Aksit, M. (2022). Comparison of hierarchic clustering methods with cophenetic correlation coefficient in big data. *Afyon Kocatepe University Journal of Sciences and Engineering*, 22, 552–559. <https://doi.org/10.35414/akufemubid.1018302>
- Saracli, S., Dogan, N., & Dogan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013, 1–8. <https://doi.org/10.1186/1029-242X-2013-203>
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, 664–681. <https://doi.org/10.1016/j.neucom.2017.06.053>
- Schober, P., & Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia*, 126(5), 1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>
- Scoltock, J. (1982). A survey of the literature of cluster analysis. *Computer Journal*, 25(1), 130–134. <https://doi.org/10.1093/comjnl/25.1.130>
- Sedgwick, P. (2014). Spearman's rank correlation coefficient. *BMJ (Online)*, 349, 1–3. <https://doi.org/10.1136/bmj.g7327>
- Shafidan, N. A. J., & Sharif, S. (2023). Soft Skills Segmentation and Profiling Analysis. *Journal of Islamic, Social, Economics and Development (JISED)*, 8(56), 275–290. <https://doi.org/10.55573/JISED.085624>
- Sharif, S., Yusoff, N. S., & Djauhari, M. A. (2012). Network topology of foreign exchange rate. *Modern Applied Science*, 6(11), 35–43. <https://doi.org/10.5539/mas.v6n11p35>
- Shetty, P., & Singh, S. (2021). Hierarchical Clustering: A Survey. *International*

- Shirkhorshidi, A. S., Aghabozorgi, S., & Ying Wah, T. (2015). A Comparison study on similarity and dissimilarity measures in clustering continuous data. In *PLoS ONE* (Vol. 10, Issue 12). <https://doi.org/10.1371/journal.pone.0144059>
- Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling and Software*, 22(4), 464–475. <https://doi.org/10.1016/j.envsoft.2006.02.001>
- Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) - A case study. *Water Research*, 38(18), 3980–3992. <https://doi.org/10.1016/j.watres.2004.06.011>
- Singh, V. K., & Tewari, D. V. (2022). Design of Improved 3D Radar charts for Multidimensional Data Visualization. *International Journal of Emerging Trends in Engineering Research*, 10(12), 455–460. <https://doi.org/10.30534/ijeter/2022/0110122022>
- Sokal, R. R., & Rohlf, F. J. (1962). The Comparison of Dendrograms By Objective Methods. *Taxon*, 11(2), 33–40. <https://doi.org/10.2307/1217208>
- Sokal, R.R & Michener. C. D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Scientific Bulletin*, 28, 1409–1438.
- Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), 72–101.
- Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. *New Directions in Statistical Physics*, 273–309. https://doi.org/10.1007/978-3-662-08968-2_16
- Stooksbury, D. E., & Michaels, P. J. (1991). *Cluster Analysis of Southeastern U.S. Climate Stations*. 150, 143–150.
- Strauss, T., & Von Maltitz, M. J. (2017). Generalising ward’s method for use with manhattan distances. In *PLoS ONE* (Vol. 12, Issue 1, pp. 1–21). <https://doi.org/10.1371/journal.pone.0168288>
- Strickert, M., Sreenivasulu, N., Villmann, T., & Hammer, B. (2008). Robust centroid-based clustering using derivatives of pearson correlation. *BIOSIGNALS 2008 - International Conference on Bio-Inspired Systems and Signal Processing*, 2, 197–203. <https://doi.org/10.5220/0001062601970203>
- Sun, L., Chen, G., Xiong, H., & Guo, C. (2017). Cluster Analysis in Data-Driven Management and Decisions. *Journal of Management Science and Engineering*, 2(4), 227–251. <https://doi.org/10.3724/sp.j.1383.204011>
- Szekely, G. J., & Rizzo, M. L. (2005). Hierarchical Clustering via Joint Between-

- Within Distances: Extending Ward's Minimum Variance Method. In *Journal of classification* (Vol. 22, Issue 2, pp. 151–183).
- Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. In *Journal of Diagnostic Medical Sonography* (Vol. 6, Issue 1, pp. 35–39). <https://doi.org/10.1177/875647939000600106>
- Temizhan, E., Mirtagioglu, H., & Mendes, M. (2022). Which Correlation Coefficient Should Be Used for Investigating Relations between Quantitative Variables? *American Academic Scientific Research Journal for Engineering, Technology, and Sciences*, 85(1), 265–277. https://www.asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/7326/2499
- Tilevik, A. (2017). *Evaluation of clustering methods for analyzing drug cytokine profiles*.
- Tokuda, E. K., Comin, C. H., & Costa, L. da F. (2020). Revisiting Agglomerative Clustering. In *Physica A: Statistical Mechanics and its Applications* (pp. 1–22). <https://doi.org/10.1016/j.physa.2021.126433>
- Toriman, M. E., Hashim, N., Hassan, A. J., Mokhtar, M., Juahir, H., Gasim, M. B., & Abdullah, M. P. (2011). Study on the Impact of Tidal Effects on Water Quality Modelling of Juru River, Malaysia. *Asian Journal of Scientific Research*. <https://doi.org/10.3923/ajsr.2011.129.138>
- Torres, G. J., Basnet, R. B., Sung, A. H., Mukkamala, S., & Ribeiro, B. M. (2009). A Similarity Measure for Clustering and its Applications. *Direct*, 31(July 2008), 490–496. <https://org.coloradomesa.edu/~rbasnet/research/ClusteringSimilarityAndItsApplications.pdf>
- Tuğran, E., Kocak, M., Mirtagioglu, H., Yiğit, S., & Mendes, M. (2015). A Simulation Based Comparison of Correlation Coefficients with Regard to Type I Error Rate and Power. *Journal of Data Analysis and Information Processing*, 03(03), 87–101. <https://doi.org/10.4236/jdaip.2015.33010>
- Ugwu, S., Miasnikof, P., & Lawryshyn, Y. (2023). Distance Correlation Market Graph: The Case of S&P500 Stocks. *Mathematics*, 11(3832), 1–13. <https://doi.org/10.3390/math11183832>
- Vidman, L., Kallberg, D., & Ryden, P. (2019). Cluster analysis on high dimensional RNA-seq data with applications to cancer research - An evaluation study. *PLoS ONE*, 14(12), 1–21. <https://doi.org/10.1371/journal.pone.0219102>
- Vijaya, Aayushi, S., & Bateja, R. (2017). A Review on Hierarchical Clustering Algorithms. In *Journal of Engineering and Applied Sciences* (Vol. 12, Issue 24, pp. 7501–7507).
- Vijaya, V., Sharma, S., & Batra, N. (2019). Comparative Study of Single Linkage, Complete Linkage, and Ward Method of Agglomerative Clustering. *International Conference on Machine Learning, Big Data, Cloud and Parallel*

- Vukotić, S., & Mirčetić, V. (2020). Clustering in Agriculture and Tourism as a Potential for Development of Rural Tourism. *Tourism in Function of Development of The Republic of Serbia Tourism and Rural Development (TISC 2020) – Thematic Proceedings II CLUSTERING*, 470–487.
- Wang, H., Wang, W., Yang, J., & Yu, P. S. (2002). Clustering by pattern similarity in large data sets. *Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data, SIGMOD 2002*, 2, 394–405. <https://doi.org/10.1145/564691.564737>
- Wang, X., & Xu, Y. (2019). An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. *IOP Conference Series: Materials Science and Engineering*, 569, 1–6. <https://doi.org/10.1088/1757-899X/569/5/052024>
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. In *Journal of the American Statistical Association* (Vol. 58, Issue 301, pp. 236–244). <https://doi.org/10.1080/01621459.1963.10500845>
- Warsito, B., Sumiyati, S., Yasin, H., & Faridah, H. (2021). Evaluation of river water quality by using hierarchical clustering analysis. *IOP Conference Series: Earth and Environmental Science*, 896(1). <https://doi.org/10.1088/1755-1315/896/1/012072>
- Wiedermann, W., & Hagmann, M. (2016). Asymmetric properties of the Pearson correlation coefficient: Correlation as the negative association between linear regression residuals. *Communications in Statistics - Theory and Methods*, 45(21), 6263–6283. <https://doi.org/10.1080/03610926.2014.960582>
- Wijuniamurti, S., Nugroho, S., & Rachmawati, R. (2022). Agglomerative Nesting (AGNES) Method and Divisive Analysis (DIANA) Method For Hierarchical Clustering On Some Distance Measurement Concepts. *Journal of Statistics and Data Science*, 1(1), 7–11. <https://doi.org/10.33369/jsds.v1i1.21009>
- Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. *British Journal of Mathematical & Statistical Psychology*, 46, 339–349.
- Wilcox, R. R. (1994). Estimating Winsorized correlations in a univariate or bivariate random effects model. *British Journal of Mathematical and Statistical Psychology*, 47(1), 167–183. <https://doi.org/10.1111/j.2044-8317.1994.tb01030.x>
- Wilcox, R. R. (2004). Inferences based on a skipped correlation coefficient. *Journal of Applied Statistics*, 31(2), 131–143. <https://doi.org/10.1080/0266476032000148821>
- Wilcox, R. R. (2016). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, 1–10. <https://doi.org/10.1111/bmsp.12069>

- Wu, S., & Chow, T. W. S. (2003). Self-organizing-map based clustering using a local clustering validity index. *Neural Processing Letters*, 17(3), 253–271. <https://doi.org/10.1023/A:1026083612746>
- Xao, O. G., Yahaya, S. S. S., Abdullah, S., & Yusof, Z. M. (2014). H-statistic with winsorized modified one-step M-estimator for two independent groups design.pdf. *AIP Conference Proceedings* 1605, 928–931.
- Xu, D., & Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2(2), 165–193. <https://doi.org/10.1007/s40745-015-0040-1>
- Xu, N., Finkelman, R. B., Dai, S., Xu, C., & Peng, M. (2021). Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal. *ACS Omega*, 6, 6206–6217. <https://doi.org/10.1021/acsomega.0c05758>
- Xu, N., Peng, M., Li, Q., & Xu, C. (2020). Towards consistent interpretations of coal geochemistry data on whole-coal versus ash bases through machine learning. *Minerals*, 10(4). <https://doi.org/10.3390/min10040328>
- Xu, R., & Wunsch, D. C. (2010). Clustering algorithms in biomedical research: A review. *IEEE Reviews in Biomedical Engineering*, 3, 120–154. <https://doi.org/10.1109/RBME.2010.2083647>
- Yim, O., & Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology*, 11(1), 8–21. <https://doi.org/10.20982/tqmp.11.1.p008>
- Yusniyanti, A., Virgantari, F., & Faridhan, Y. (2021). Comparisan of Average Linkage and K-Means Methods in Clustering Indonesia's Provinces Based on Welfare Indicators. In *Journal of Physics: Conference Series* (Vol. 161, Issue 1, pp. 0–10).
- Zaini, B. J., & Sharif, S. (2020). The performance of different correlation coefficient under contaminated bivariate data. *Mathematics and Statistics*, 8(2), 1–8. <https://doi.org/10.13189/ms.2020.081301>
- Zainudin, E. Z. (2010). Benchmarking river water quality in Malaysia. *Jurutera*, February, 12–15.
- Zakaria, N. A., Abdullah, S., Ahad, N. A., Yusof, N., & Yahaya, S. S. S. (2016). The performance of robust correlation coefficient under contaminated bivariate data. *American Institute of Physics*. <https://doi.org/10.1063/1.4966107>
- Zar, J. H. (2005). Spearman Rank Correlation. *Encyclopedia of Biostatistics*, 1–9. <https://doi.org/10.1002/0470011815.b2a15150>
- Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 25(2), 103–114.

<https://doi.org/10.1145/235968.233324>

Zhou, S., Xu, Z., & Liu, F. (2017). Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 28(12), 3007–3017. <https://doi.org/10.1109/TNNLS.2016.2608001>

Zinzendoff-Okwonu, F., Laro Asaju, B., & Irimisose Arunaye, F. (2020). Breakdown Analysis of Pearson Correlation Coefficient and Robust Correlation Methods. *IOP Conference Series: Materials Science and Engineering*, 917(1). <https://doi.org/10.1088/1757-899X/917/1/012065>

Zulkepli, N. F. S., Md Noorani, M. S., Razak, F. A., Ismail, M., & Alias, M. A. (2020). Cluster analysis of haze episodes based on topological features. *Sustainability (Switzerland)*, 12(10). <https://doi.org/10.3390/SU12103985>

