12-2021

Functional Mixed Data Clustering with Fourier Basis Smoothing

Ishmael Amartey
East Tennessee State University

Functional Mixed Data Clustering with Fourier Basis Smoothing

---

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Ishmael Amartey

December 2021

JeanMarie Hendrickson, Ph.D., Chair

Michele Joyner, Ph.D.

Robert Price, Ph.D.

Keywords: hierarchical clustering, mixed data, Gower coefficient, functional data

ABSTRACT

Functional Mixed Data Clustering with Fourier Basis Smoothing

by

Ishmael Amartey

Clustering is an important analytical technique that has proven to affect human life positively through its application in cancer research, market segmentation, city planning etc. In this time of growing technological systems, mixed data has seen another face of longitudinal, directional and functional attributes which is worth paying attention to and analyzing. Previous research works on clustering relied largely on the inverse weight technique and B-spline in smoothing data and assessing the performance of various clustering algorithms. In 1971, Gower proposed a method of clustering for mixed variable types which has been extended to include functional and directional variables by Hendrickson (2014). In this study, we will do a comparative analysis of the performance of the hierarchical clustering mechanism using a simulated Functional data with mixed structure. We will adopt the Fourier basis smoothing procedure and use the Rand index (Rand 1971) and adjusted Rand index for the comparison of the various clustering algorithms.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. JeanMarie Hendrickson for her supervision, corrections and support given to me throughout this project. I am also grateful to Dr Robert Price and Dr Michele Joyner for accepting to be part of my committee. I'd like to thank Dr Robert Gardner for his immense support, guidance and encouragement throughout my study at ETSU. To my family and friends, I say a big thank you. Thank you Mom for being there whenever I needed you and for pushing me to do exploits. I appreciate all the sacrifices you made to make sure I come this far.

Lastly, I would like to specially thank my wife. Faith, if not for your love and support this journey would have been extra difficult. Thank you for the joy you bring to my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Over the years data acquisition has experienced tremendous improvement due to technological advancements. Now, what would have been a tiring and time-consuming procedure to gather data has been reduced to significantly lower levels with just a single click on a computer through emails and internet participation. This new advanced way of getting data has largely reduced cost but like any other process this has its own drawbacks. Some of these drawbacks include non-participation, non-response or missing data, amongst others. Data mining has become a strong component affecting every sphere of life as it is the basic source of a majority of decision making in the world at large. Because of the significance of data, in medicine, real estate, media, education, policy making et. al., it is prudent that data and its characteristics are understood to unearth the unknown. Data comes in different forms and types and has been traditionally classified under two main forms, quantitative or categorical, but within these two general classifications, there might be some unique characteristics within data sets which, when ignored, can misinform an analyst.

Hendrickson [18] acknowledged that most real data have different characteristics and variables. As a result, data types cannot be limited to the traditional two types, especially when one is dealing with mixed data. So, in order to study the unique characteristics of groups in data, clustering must be adopted. Mixed data is one that comprises of both categorical (color, sex, blood group) and quantitative (height, age, weight).

According to Chapman and Hall/CRC [15], Aristotle's classification of living and non-living things constituted the first known clustering. Aristotle classified animals

into two main groups, vertebrates and invertebrates and went further to classify how these animals reproduce [31]. In medicine, the collection of such data and its analysis is used in cancer research, vaccinations adaptation and the creation of life tables in survival analysis for cohort groups [15]. In business, these data inform decision making such as which stocks an investor must invest in to maximize profit or diversify portfolio shocks and market segmentation [16].

In modern technologies, facial and pattern recognition are widely used to provide robust security systems and data protection for users [17]. Also, search engines and social media platforms use a similarity matrix to continually suggest content to users of the internet depending on the searches they make [30]. These platforms collect data on the interest of users to make accurate group suggestions.

Other applications of clustering includes determining temperaments [27] in behavioural science and soil type in agriculture [25]. There are different types of clustering methods which can be adopted depending on the data set to be used for research. The various kinds of clustering will be discussed in detail in chapter two.

2 LITERATURE REVIEW

2.1 Introduction to Clustering

Clustering is way of determining groups in multivariate data. This is done by identifying data with similar characteristics and grouping them. With clustering, a large data set becomes convenient to work with and can be summarized and understood easily. What needs to be checked for a good clustering is that data within a cluster should be homogeneous (i.e highly similar in attributes) and data between clusters should be heterogeneous or highly dissimilar. There are different techniques in clustering data, these includes the hierarchical method, k-means method, model based method and centroid based clustering [31]. Figure (2.7) shows the various techniques in clustering data.

In hierarchical clustering, data sets are not subdivided into a certain number of clusters in a single step. Rather, the classifications are done in series of partitions from a cluster containing all possible individuals to n clusters containing just a single individual [31]. The hierarchical clustering technique can be categorized further into agglomerative method, which is continued by successive fusions of the n individuals into groups, and divisive methods which categorizes the individuals into successive distinct groups [31]. How hierarchical clustering works is by finding the least distance between data points and grouping them to form a cluster. Considering two points on a plane with each point being a cluster on its own, the measurement of the least distance between each point can be calculated using the Euclidean distance measure, the squared Euclidean distance measure, the Manhattan distance measure or the

Cosine distance measure [29]. The most commonly used distance measure is the Euclidean distance which is given by

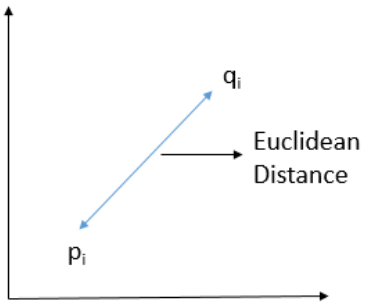

Figure 2.1: Euclidean distance

$$d = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{2.1}$$

where p_i and q_i are the Euclidean vectors starting from the origin of the space (initial point) or

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_p - y_p)^2} \tag{2.2}$$

where x and y, are the points in the euclidean $n-$ space. To further understand how this works, we will use the data from Kaufmann and Rousseeuw [21] in Table 2.1.

Table 2.1: Data from Kaufmann and Rousseeuw [21] of Seven Measured Objects for Two Variables.

Object	Variable 1	Variable 2
1	2.00	2.00
2	5.50	4.00
3	5.00	5.00
4	1.50	2.50
5	1.50	1.00
6	1.00	5.00
7	7.00	6.50

The dissimilarity between the objects can be obtained as follows. For objects 1 and 2 the squared difference between 5.50 and 2.00 for variable 1 is 12.25 and the squared difference between 2 and 4 for variable 2 is 4. Using Equation (2.2) we get $d(1,2) = \sqrt{12.25 + 4} = 4.03$. Similarly, the squared differences between objects 1 and 3 for variable 1 is 9 and that of variable 3 is also 9. Again using Equation (2.2) we get the dissimilarity between object 1 and 3 for variable 1 and variable 2 as $d(1,3) = \sqrt{9+9} = 4.24$. We continue this procedure to obtain the dissimilarity between the objects combinations $(1,4),(1,5),(1,6),(1,7),(2,3),...,(6,7)$. The dissimilarity between objects of the same kind is zero, so its suffices that the diagonal elements of the matrix are zero. The dissimilarity matrix for the seven objects from Kaufmann and Rousseeuw [21] is given by

$$\begin{bmatrix} 0.0 & & & & & & \\ 4.03 & 0.0 & & & & & \\ 4.24 & 1.12 & 0.0 & & & & \\ 0.71 & 4.27 & 4.30 & 0.0 & & & \\ 1.41 & 5.41 & 5.66 & 1.58 & 0.0 & & \\ 5.83 & 1.80 & 2.00 & 6.04 & 7.21 & 0.0 & \\ 5.86 & 2.51 & 1.68 & 5.84 & 7.27 & 1.95 & 0.0 \end{bmatrix}$$

Figure (2.2) depicts how the data points from Kaufmann and Rousseeuw [21] were merged together to form the cluster dendogram in Figure (2.3). To start the merging process, we first do a scatter plot of the data and combine closest points to form a

cluster. In Figure (2.2), objects 1 and 4 were joined to form a cluster. Same was done for objects 2 and 3, and objects 6 and 7. After this process we locate the nearest data point to the formed clusters and merge them like we did in the previous process to form another cluster. In Figure (2.2), we merged objects 5 to that of the cluster containing objects 1 and 4 to form a new cluster and merged the two clusters containing objects 2 and 3, and objects 6 and 7 to form another single cluster. We stop the merging process when there are no more data points to add to a cluster. Then, we merge all the created clusters to form one cluster containing all the objects. Figure (2.3) depicts the final hierarchical cluster dendogram after the merging process. Here, we see that objects 1 and 4 are much more similar to each other than they are to object 5. Same applies to objects 2 and 3, and objects 6 and 7. Though the dendogram is a good way to visualize the distance between objects, it can lead to loss of information. For instance in Figure (2.3), it appears the distance from object 4 to 2 is shorter than the distance from object 4 to 5 but this is false and it can clearly be noticed in the scatter plot of Figure (2.2).

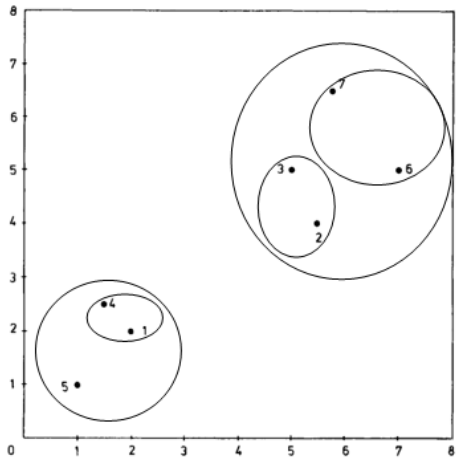

Figure 2.2: Cluster dendogram set up for the seven measured objects

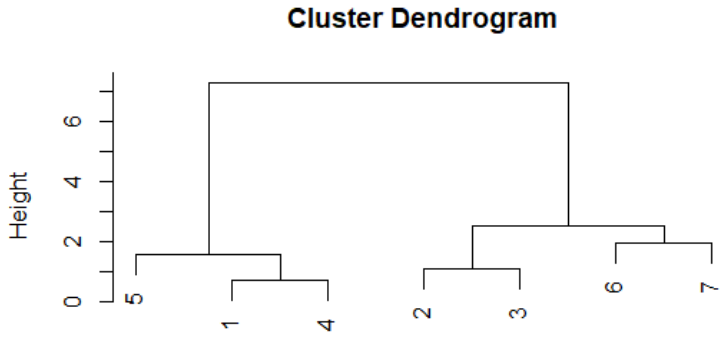

Figure 2.3: Cluster dendogram for the seven measured objects from Kaufmann and Rousseeuw

With the Manhattan distance, the measurement of the distance between two points along different axes at right angles is the point of interest, more formally it is written as

$$\boldsymbol{d} = \sum_{i=1}^{n} |q_i - p_i| \tag{2.3}$$

Figure 2.4 depicts the graph of the Manhattan distance.

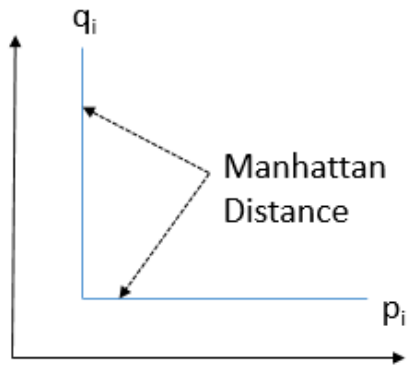

Figure 2.4: Manhattan distance

Figure 2.5 is the cosine distance between two vectors p_i and q_i. As the two vectors get further apart the cosine distance gets larger. The cosine distance is given as

$$d = \frac{\sum_{i=0}^{n-1}(q_i - p_i)}{\sum_{i=0}^{n-1}(q_i)^2 \sum_{i=0}^{n-1}(p_i)^2} \tag{2.4}$$

.

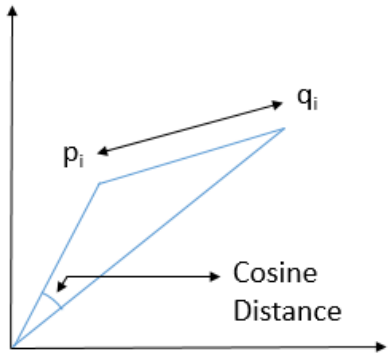

Figure 2.5: Cosine distance

Other measure of distances include the Minkowski measure of order g defined as

$$d(x,y) = (|x_1 - y_1|^g + |x_2 - y_2|^g + |x_3 - y_3|^g + ... + |x_i - y_i|^g)^{\frac{1}{g}}, \tag{2.5}$$

the Canberra metric which is define as

$$d(x,y) = \sum_{i=0}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}, \tag{2.6}$$

and the Kulczynski distance

$$\boldsymbol{d}(x,y) = \frac{\sum_{i=0}^{n} |x_i - y_i|}{\sum_{i=0}^{n} min(x_i, y_i)} \tag{2.7}$$

The $k-$means clustering method groups data into k groups. This procedure follows an easy way to classify a given data set into k clusters [20]. The main objective is to set up centroids for each cluster by partitioning them in a well structured manner to get accurate results. This is important because placing a centriod in a different lo-

cation can affect the clustering outcome [20], so the best way to begin is to randomly place centroids far away from the given data points. In Figure 2.6(A) the randomly placed centroid is denoted with (★). Next is to determine the distance from each data point to the upper and lower centroid (★) and place each data point in a lower or upper group. For instance, if the distance between a point and the upper assigned centroid is shorter than the distance between the same point and the lower centroid, then that point will be placed above the Euclidean line to form part of the upper group and vice versa. Next is to find the centroid (center) of the formed groups and continue placing points in each group using their distances from the centroid using the same approach above. In Figure 2.6(B) we assigned (▲) as the new centroid and in Figure2.6(C) we assigned (■). This process is continued until the data points converge, i.e there is no overlapping of points into another group.

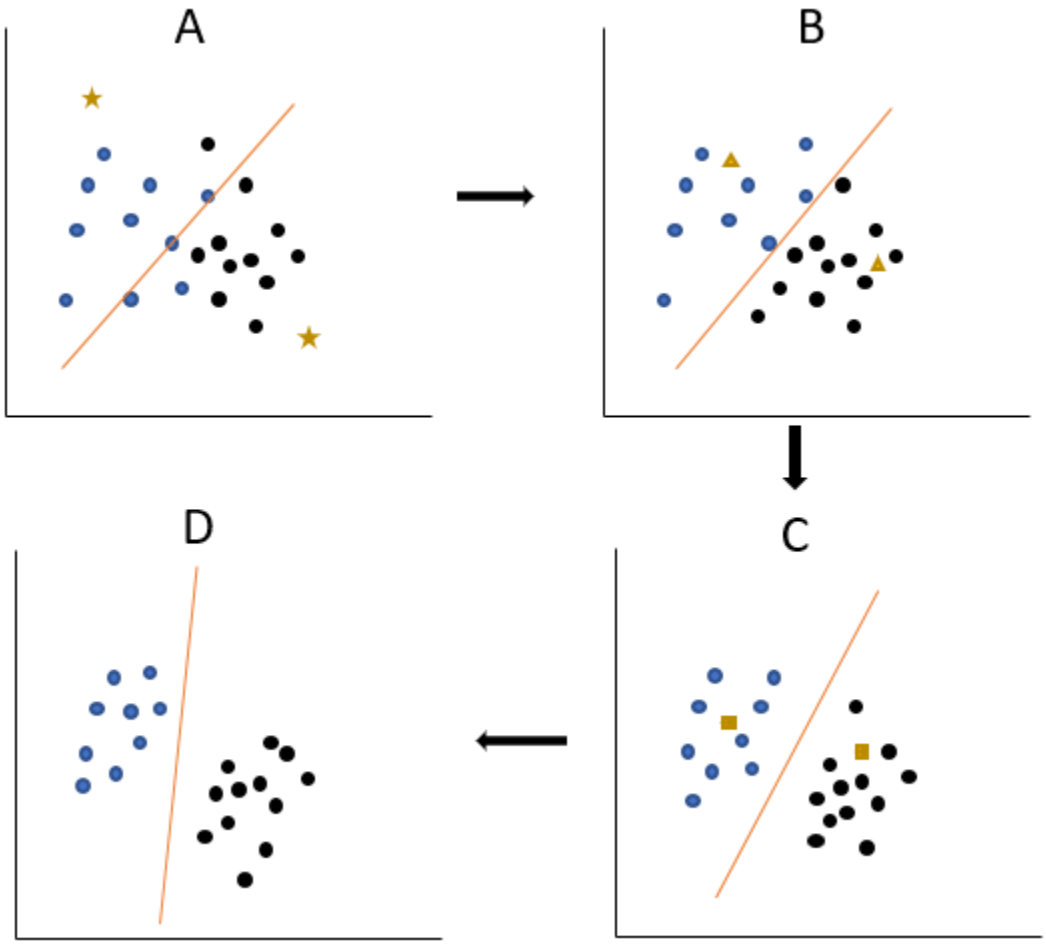

Figure 2.6: K-Means algorithm steps

In $k-$means method, the aim is to minimize the error function given as

$$W(S, C) = \sum_{k=1}^{n} \sum_{i \in s_k} ||y_i - c_k||^2, \tag{2.8}$$

where S is a $k-$cluster partition which is represented by vectors $y_i (i \in I)$, S_k is a non-overlapping cluster with a centroid c_k within.

The model-based clustering method is an alternative to the $k-$means method. It

comes with the assumption that the data comes from a distribution that is made up of two or more clusters. Unlike the above mentioned clustering methods, the model-based method uses probabilistic distributions to create clusters using the Gaussian distributions with their mean and covariance. Fraley and Raftery [12] did an extensive work on model-based clustering; we will discuss that in detail in section 2.3 of this chapter. Figure (2.7) shows the various types of clustering algorithms used in clustering.

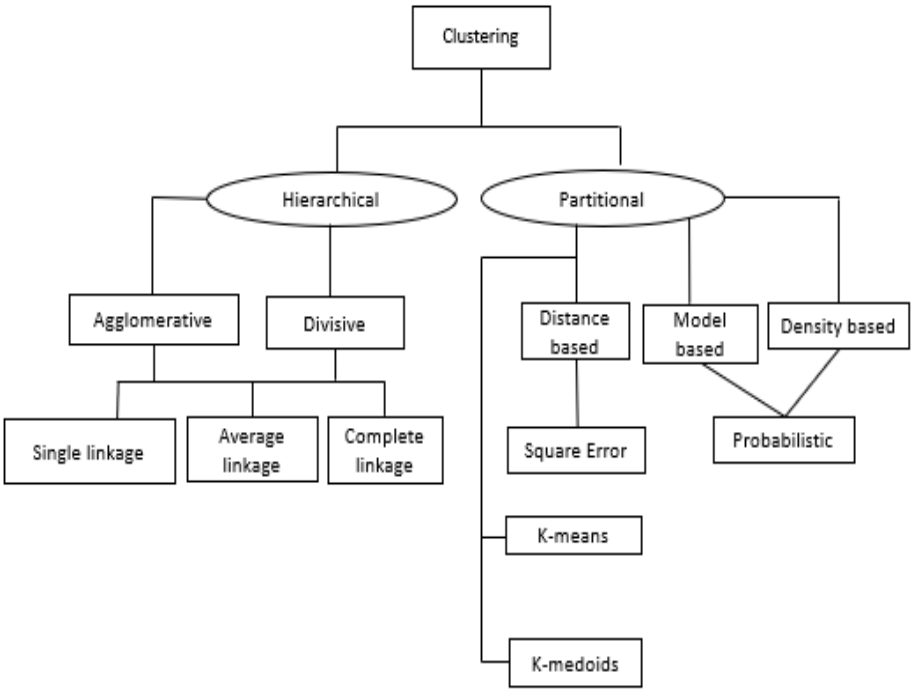

Figure 2.7: Clustering algorithms

2.2 Binary, Nominal and Ordinal variables

There are several approaches to clustering mixed data depending on the nature of data points. For instance in 2018, M.V et al clustered three mixed data sets and concluded that there is not a single cluster method that is absolute for all data sets [24]. In most cases variables are of the form binary, nominal, ordinal and interval or a combination of two or more. When a data set is binary, it is usually assigned a number value of 1 or 0 and can be measured using a contingency table.

In Table (2.2) a is the number of variables for which the assigned binary value of 1 was recorded for object 1 and object 2 and b being the number of variables for

Table 2.2: Contingency Table for Binary Data

		Object 2	
		1	0
Object 1	1	a	b
	0	c	d

which an object combination is (1,0), (0,1) for c and (0,0) for d. Binary variables can be symmetric (for instance male or female) where equal weight is assigned with no preference given over the other or asymmetric (if its states are not equally important). Zubin, J [36] adopted a simple matching coefficient method to measure the similarity in the way people behave. The simple matching coefficient for dissimilar objects in binary data is given as

$$d(x, y) = \frac{b + c}{a + b + c + d} \tag{2.9}$$

When binary objects are similar their similarity is given as

$$s(x, y) = \frac{a + d}{a + b + c + d} \tag{2.10}$$

This is called the Jaccard coefficient [26]. In cases where there are more than two states, the data becomes nominal. For instance, when one is studying the nationality of people, each country can be coded as 1,2,3,...,M where M denotes the total number of states. It should be noted that the assigned values are only for coding purposes, and that the states are not ordered in any particular way. Sometimes nominal variables can be broken down to only two states to form a binary variable, but this procedure can cause loss of information [25].

The simple matching approach can be used to measure the similarity or dissimilar-

ity between objects that takes the form of a nominal variable. The similarity measure for such objects is defined as

$$s(x, y) = \frac{u}{p} \tag{2.11}$$

and the dissimilarity is defined as

$$d(x, y) = \frac{p - u}{p} \tag{2.12}$$

where $u = a + d$, the number of matches for which the objects are in the same state and $p = a + b + c + d$, the total number of variables as defined in Table 2.2 [33].

Unlike nominal variables, ordinal variables place emphasis on the asssigned values or states. For instance, in a study to determine how satisfied customers are after using a certain product the scale values could be: 1-Very dissatisfied, 2- Dissatisfied, 3- Fair, 4- Satisfied and 5-Very satisfied. For such ordinal variables we map the range between 0 and 1 such that each variable achieves an equal weighing. This is defined as

$$z_{in} = \frac{r_{in} - 1}{M_n - 1} \tag{2.13}$$

where z_{in} is the standardized value for object i in variable n, r_{in} is the rank of the ith object in the nth variable and M_n is the highest rank for variable n [21].

2.3 Review on Mixed Variables

While binary, ordinal, and nominal variables are very common, in real life applications its very common that several kinds of variables are present in the data set. Table

(2.3) is a table of the characteristics of a garden flower (Taken from Kauffman and Rousseeuw, 1990.pg 33) where W represents winter (Yes=1, No=0), S for shadow (Yes=1, No=0), T for tubers (Yes=1, No=0), Col for colour of flowers (White=1, Yellow=2, Pink=3, Red=4, Blue=5), Soil (Dry=1, Normal=2, Humid=3), Pr for preference (Low=1, High=18), H is height in centimeters and PD is planting distance in centimeters.

Table 2.3: Characteristics of Some Garden Flowers

	Garden Flower	W	S	T	Col	Soil	Pr	H	PD
1	Begonia (Bertinii bolivieness)	0	1	1	4	3	15	25	15
2	Broom (Cytisus praecox)	1	0	0	2	1	3	150	50
3	Carnellia (Japonica)	0	1	0	3	3	1	150	50
4	Dahlia (Tartini)	0	0	1	4	2	16	125	50
5	Forget-me-Not (Myosotis sylvatica)	0	1	0	5	2	2	20	25
6	Fuchsia (Marinka)	0	1	0	4	3	12	50	40
7	Geranium (Rubin)	0	0	0	4	3	12	50	40
8	Gladiolus (Flowersong)	0	0	1	2	2	7	100	15
9	Heather (Erica carnea)	1	1	0	3	1	4	25	15
10	Hydrangea (Hortensis)	1	1	0	5	2	14	100	60
11	Iris (Versicolor)	1	1	1	5	3	8	45	10
12	Lily (Lilium regale)	1	1	1	1	2	9	90	25
13	Lily-of-the-valley (Convallaria)	1	1	0	1	2	6	20	10
14	Peony (Paeonia lactiflora)	1	1	1	4	2	11	80	30
15	Pink Carnation (Dianthus)	1	0	0	3	2	10	40	20
16	Red Rose (Rosa rugosa)	1	0	0	4	2	18	200	60
17	Scotch Rose (Rossa pimpinella)	1	0	0	2	2	17	150	60
18	Tulip (Tulipia sylvestris)	0	0	1	2	1	5	25	10

In dealing with mixed variables one can treat each variable as a single cluster rather than mixing them, this procedure is only accepted when the conclusions from the single variables agree. The drawback of treating each variable as a cluster is that when different results are obtained it becomes difficult to reconcile them [21]. So it is prudent to treat the mixed data together and proceed to do a single cluster analysis.

The focus on clustering analysis has largely been on finding the dissimilarity between mixed data variables, but Gower [14] proposed a coefficient to find the similarity

between mixed data. This is of the form

$$\boldsymbol{S}_{ij} = \frac{\sum_{t=1}^{p} \boldsymbol{s}_{ijt}\delta_{ijt}}{\sum_{ijt}^{p} \delta_{ijt}} \tag{2.14}$$

Gower assigned weights where δ_{ijt} represent when there is possibility in comparisons between the individual characters i and j at t and assigned scores for s_{ijt} as follows: i) For characters with exactly two outcomes, a presence of a character is assigned + and - otherwise. When there is an unknown value, a $2X2$ contingency table is used to assign weights as shown in the table below.

Table 2.4: Scores for Characters with Two Outcomes

	Values of t
Individual i Individual j	$++--$ $+-+-$
s_{ijt} δ_{ijt}	1 0 0 0 1 1 1 0

ii) In the case of a qualitative character

$$s_{ijt} = \left\{ \begin{array}{ll} 1, & \text{if there is an agreement between i and } j \text{ at } t \\ 0, & \text{otherwise} \end{array} \right\}$$

iii) When characters are quantitative

$$s_{ijt} = 1 - \frac{|x_i - y_j|}{R_t} \tag{2.15}$$

where R_t is the range of t. Kaufmann and Rousseeuw [21] later generalalized the Gower's distance as the complement of the similarity coefficient proposed by Gower

in 1971 as

$$d(i,j) = 1 - s_{ij} = \frac{\sum_{t=1}^{p} s_{ijt}\delta_{ijt}}{\sum_{ijt}^{p} \delta_{ijt}} \tag{2.16}$$

The Gower coefficient came with a drawback of making one variable dominant over the other since it assigned equal weights to either of the variable types whether its continuous or binary. In 2006, Chae, Kim, and Yang [19] rectified the draw back from Gower's coefficient by assigning different weights to different variable types. They defined the dissimilarity measure as:

$$d_{ij}^{*} = \tau_{ij} \sum_{l=1}^{c} \frac{1}{C} \left(\frac{|x_{il} - y_{jl}|}{R_l} \right) + (1 - \tau_{ij}) \sqrt{1 - \frac{\sum_{l=c+1}^{r} s_{ijl}}{\sum_{l=c+1}^{r} w_{ijl}}} \tag{2.17}$$

where p_{ij}^{c} is the Pearson correlation coefficient for the quantitative variable, p_{ij}^{d} is the product moment correlation for multiple binary variables and $\{\tau_{ij} : 0 \leq \tau_{ij} \leq 1\}$ is a balancing weight to prevent dominance of one attribute over the other satisfying

$$\tau_{ij} = \left\{ \begin{array}{cc} 1.0 - \frac{p_{ij}^{c}}{|p_{ij}^{c}| + |p_{ij}^{d}|}, & \text{if } 1.0 < \frac{p_{ij}^{c}}{p_{ij}^{d}} \\ 1.0 - \frac{p_{ij}^{d}}{|p_{ij}^{c}| + |p_{ij}^{d}|}, & \text{if } 1.0 > \frac{p_{ij}^{c}}{p_{ij}^{d}} \\ 0.5, & \text{if } |p_{ij}^{c}| = |p_{ij}^{d}| \end{array} \right\}$$

with $-1.0 \leq p_{ij}^{c}$ as the similarity measure for the quantitative variable, $p_{ij}^{d} \leq 1.00$ as the measure of similarity for the binary variables, $i = 2, 3, ..., n.$ and $j = 1, 2, ..., n-1$ for $i > j$. R_l is the range of the l_{th} variable in quantitative values and $w_{ijl} = 1$ for continuous variables, $s_{ijl} = 0$ if $x_i = y_j$ and 0 otherwise, for binary variables. w_{ijl} could take the value of 0 or 1 for binary variables provided there is a valid comparison between the i_{th} and j_{th} objects for variables in the l_{th} position.

In the study of Chae, Kim and Yang [19], they adopted the use of the correlation coefficient (p_{ij}^c) for the quantitative variables and the product moment correlation (p_{ij}^d) for the case of multiple binary variables. However, they indicated that any reasonable measure of similarity between the i_{th} and $_{jth}$ objects within different kinds of variables could be used and not necessarily p_{ij}^c and p_{ij}^d.

Clustering with mixed data can be complex. However, a model-based method produces reasonably good partitions without prior information about the data groupings [12]. Using a the Bayesian Information criterion (BIC), Schwartz [32] determined the number of groups in a data set by initializing the expectation-maximization (EM) with partitions from a model-based algorithm. Following that a good outcome from Dasgupta and Raftery [5] on minefield and seismic fault detection, Fraley and Raftery [11] extended to select clusters simultaneously with the use of the BIC. The BIC is of the form

$$2\log p(D|M_{\mathrm{k}}) \approx 2\log p(D|\hat{\theta}_k, M_k - v_k \log(n)) = BIC_k \tag{2.18}$$

where v_k is the number of independent parameters to be estimated in the model M_k, $p(D|M_k)$ is the integrated likelihood of model M_k defined as

$$p(D|M_k) = \int p(D|\theta_k, M_k)p(\theta_k|M_k)d\theta_k \tag{2.19}$$

where $p(D|\theta_k, M_k)$ is the prior distribution of θ_k. When independent multivariate observations are present in the data, the likelihood for such a mixture model with G

components is

$$\mathcal{L}_{\text{MIX}}(\theta_1, ..., \theta_G : \tau_1, ..., \tau_G|y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k) \tag{2.20}$$

where "f_{k} and θ are the density parameters of the kth component" as stated by Fraley [12] with $\tau \geqslant 0; \sum_{k=1}^{G} \tau_k = 1$ and f_k usually being the Gaussian normal distribution with parameters mean(μ_k) and covariance matrix $\sum_k$. The Gaussian normal distribution is defined as

$$\theta(y_i|\mu_k, \sum_k) \equiv \frac{\exp\{-\frac{1}{2}(y_i - \mu_k)^\tau \sum_k^{-1}(y_i - \mu_k)\}}{\sqrt{det(2\pi \sum_k)}} \tag{2.21}$$

Another method to clustering using the model-based method is the finite mixture model [23]. Supposing that a set of random variables $X_1, X_2, ..., X_n$, are independent and identically distributed p-dimensional observations with probability density function (pdf)

$$f(x : \pi) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{2.22}$$

where π_k is the mixing proportion present in the k_{th} sub-population, with $\pi = (\pi_1, \pi_2, ..., \pi_K)\prime$ lying in the (K-1) dimensional simplex, K the total number of components with, $f_k(x)$ is the density function and $\sum_{k=1}^{K} \pi_k = 1$. With $f_k(x) \equiv f_k(x; v_k)$, (2.22) can be written in the form

$$f(x; v) = \sum_{k=1}^{K} \pi_k f_k(x; v_k) \tag{2.23}$$

where v is the parameter to be estimated and v=$(\pi\prime, v\prime_i, v\prime_2, ..., v\prime_K)\prime$. Then we say $f(x; v)$ is a finite mixture model density with parameter vector v given as $v = (\pi\prime, v_1\prime, ..., v_k\prime)$[23]. Fraley and Raftery [12] stated that finite mixture models do not conform to the fundamental regularity conditions of the proof of the BIC proposed by Schwartz [32]. However, results show that it has a good performance in model-based clustering [12].

Another form of the model-based method are the expectation maximization (EM) which uses the maximum likelihood approach on multivariate data ([6],[22]). If the set of data (x_i) is independent and identically distributed with regards to a probability function with parametization on θ and consists of k multivariate observations with both the observed (y_i) and unobserved (z_i), then the complete-data likelihood function is of the form

$$\mathcal{L}_C(x_i|\theta) = \prod_{i=1}^{k} f(x_i|\theta) \tag{2.24}$$

When the probability that a certain variable is unobserved solely depend on the observed y variables rather than z, then the observed-data likelihood, $\mathcal{L}_O(y_i|\theta)$ is generated by integrating z out of $\mathcal{L}_C(z_i|\theta)$ to get

$$\mathcal{L}_O(y|\theta) = \int \mathcal{L}_C(x|\theta)dz \tag{2.25}$$

The EM is cycled around two steps namely; the $E-$step and the $M-$step [12]. The $E-$step is conditioned on the expectation of the log-likelihood of the complete data provided the current parameter estimates and observed data is computed; whereas, the $M-$step involves the determination of parameters that maximize the expected

log-likelihood from the $E-$step. Fraley [12] stated "In a mixture model for an EM, the 'complete data' are considered to be $X_i = (y_i, z_i)$ where $z_i = (z_{1l}, ..., z_{iG})$ is the unobserved portion of the data" with z_{ik} equal to 1 whenever X_i belongs to group k and 0 otherwise. If z_i is assumed to be an independent and identically distributed variable following a multinomial distribution of one draw from G categories associated with probabilities $\tau_1, ..., \tau_G$ with the density of y_i given as z_i is of the form

$$\prod_{k=1}^{G} = f_k(y_i|\theta_k)^{zik} \tag{2.26}$$

then the associated complete log-likelihood becomes

$$l(\theta_k, \tau_k, Z_{ik}|X) = \sum_{i=1}^{n}\sum_{k=1}^{G} Z_{ik}\log[\tau_k f_k(y_i|\theta_k)] \tag{2.27}$$

The $E-$step is set up to be

$$\hat{Z}_{ik} \longleftarrow \frac{\hat{\tau}_k f_k(y_i|\hat{\theta}_k)}{\sum_{i=1}^{G}\hat{\tau}_j f_j(y_i|\hat{\theta}_j)}. \tag{2.28}$$

To get the $M-$step, equation (2.27) will be maximized in terms of the parameters τ_k and θ_k with Z_{ijk} fixed at the values computed in the $E-$step. Some limitations of the EM method include the possibility of slower rate convergence even though it gives reasonably good results. Also the EM may fail if there are few observations present in a cluster, this typically happens when there are too many components in the multivariate data.

Everitt [8] suggested a mixture model for mixed mode data emphasizing that with

every observed variable (ordinal or categorical) there is an underlying factor that gives rise to latent continuous variables. He made reference to "threshold" of values in a given set. Everitt [8] assumed a density function of the form

$$f(x) = \sum_{i=1}^{c} p_i MVN_{p+q}\left(\mu_i, \sum\right) \tag{2.29}$$

for a vector x that contains the set of random variables $x_1, x_2, ..., x_p, x_{p+1}, ..., x_{p+q}$. From the density function c represents the assumed number of clusters in the data, $p_1, ..., p_c$ are the mixing proportions such that $\sum_{i=1}^{c} p_i = 1$, $\sum$ is the covariance matrix from a $(p+q)$ dimensional multivariate normal with mean vector μ_i. To estimate the parameters of Everitt's [8] model, a maximum likelihood approach can be used with a suitable optimization algorithm. Though the model gave reasonably good results, it has a limitation of not being feasible for larger values of q, ordinal and binary variables.

3 RECENT COMPARATIVE WORK AND PROPOSED WORK

Much work on clustering has been geared towards single linear data sets with very few data sets involving mixed data. Attention to include other variable structure, particularly directional, began in 1918 by Von Mises [35] followed by other research on spherical and hyper-spherical data attributes by Fisher [10]. In 2014, Hendrickson [18] extended the Gower coefficient to cater to functional and directional data curtailing the drawbacks associated with the traditional method of converting nominal variables into numeric variables which leads to loss of information. In studies that focused on mixed data, the most popular smoothing technique has been the B-spline. Notable amongst these studies include the work of Laura Ferreira and David B. Hitchcock [9], Obed Oppong [2], Augustine Koomson [25]. Several methods of clustering have been designed for various data settings. In clustering mixed data with other variable attributes, the choice of distance calculation and assigning weight functions are key as it can improve the performance of the statistical function used in the clustering process. The most common weight function is the inverse weight function which uses the variance of the observed functional data. Chen, Reiss and Tarpey [4] proposed a new method of adding weight to functional data called the CV optimal weighing using the coefficient of variation. Tapey [34] elaborated on how different clustering results can be achieved depending on how data curves are fitted and the kind of basis function used. Tapey, asserted that clustering functional data with the L^2 metric on function produces similar results by applying a suitable clustering mechanism to a linear transformation of the regression coefficient.

3.1 Proposed Work

This study will entail the simulations of data generated in the same manner as that of Hendrickson [18] with functional and directional attributes to access the performance of the single, average and complete linkage hierarchical clustering using the Fourier basis smoothing. To access how each clustering mechanism perform, we shall use the Rand and adjusted rand index to make a determination when weights are applied to the functional data and when they are unweighted. The weighting scheme to use will be the proposed weighing technique by Chen et al [4]. We shall implore the extension of the Gower coefficient as the dissimilarity measure for the functional data in this study.

3.2 Extention of the Gower Coefficient

Hendrickson [18] extended the Gower coefficient to make room for functional and directional data attributes. For objects i and j, the dissimilarity is defined as

$$d(i,j) = \frac{\sum_f \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_f \delta_{ij}^{(f)}} \tag{3.1}$$

where $\delta_{ij}^{(f)} = 1$ if the measurements x_{if} and y_{jf} for the fth variable are non-missing and 0 otherwise [21].

- for binary or nominal values of f,

$$d_{ij}^{f} = \left\{ \begin{array}{cc} 1 & if \quad x_{if} \neq y_{jf} \\ 0 & if \quad x_{if} = y_{jf} \end{array} \right\}.$$

- If all variables are nominal or symmetric binary, then d_{ij} is equal to the matching coefficient.

To include the directional variables, Hendrickson adopted the Ackermann [1] dissimilarity measure for directional variables defined as

$$d_{ij}^{f} = \pi - |\pi - |\theta_i - \theta_j|| \tag{3.2}$$

where θ is the angle measured between variables of object i and j. Other measures of dissimilarity for this study include:

- L_1 distance for interscaled-scaled variables defined as

$$d_{L_1}^{(f)}(i,j) = |x_{if} - y_{jf}|$$

- L_2 distance for functional variables defined as

$$d_{L_2}^{(f)}(i,j) = \sqrt{\int_T w(t)[x_{if} - y_{j,f}]^2 df}$$

where $w(t) \geq 0$ is a defined weight function.

3.3 Fourier Basis

In natural settings, data comes with some sort of observational error or noise which can influence the outcome of analysis [28]. These noise tends to hide the underlying trend in the data and thus causes one to underfit or overfit the underlying trend.

A method to reduce the effects of these noise is by using a smoothing technique. The Fourier basis is one of many ways to smooth data gathered over a period of time whether equally intervaled or not [28]. In this study we shall adopt a linear combination of functions to define the functional observation $\varphi(i)$ as

$$\varphi_i(t) \approx \sum_{n=1}^{n} c_{in}\delta_n(t), \forall t \in \mathbb{T} \tag{3.3}$$

where δ_n is defined as

$$\delta_n(t) = \frac{1}{\sqrt{|\mathbb{T}|}} \tag{3.4}$$

$$\delta_{2r-1}(t) = \frac{\sin r\omega t}{\sqrt{\frac{|\mathbb{T}|}{2}}} \tag{3.5}$$

and

$$\delta_{2r}(t) = \frac{\cos r\omega t}{\sqrt{\frac{|\mathbb{T}|}{2}}} \tag{3.6}$$

for $r = 1, ..., \frac{k-1}{2}$, ω is the period and $|\mathbb{T}| = \frac{2\pi}{\omega}$

Here, t is a time variable with elements $\{t_1, t_2, ..., t_j\} \in \mathbb{T}$, δ_n for $n = (1, 2, ..., N)$ is the n^{th} basis function of the expansion and c_{in} is the associated coefficient. For an N observational data, $X = [X_1^T, ..., X_H^T]^T$, the functional data is defined as

$$X_i = \mathcal{Z}_i(t_j) + \varepsilon_i, 1 \leq j \leq J, 1 \leq i \leq N, \tag{3.7}$$

where X_i is an observation with noise as a result of the stochastic process, $\mathbb{Z}_i(t_j)$

related to the i^{th} functional data and $\varepsilon_i \sim$ random error with mean zero and variance σ_i^2. The stochastic process $\mathbb{Z}_i(t_j)$ is given as

$$\mathbb{Z}_i(t) \approx c_i^{\mathbb{T}} \delta(t), \forall t \in \mathbb{T}, i = 1, ..., N \tag{3.8}$$

where c_i and $\delta(t)$ are N vectors. The Fourier basis has a constant value as its first element and then alternates between the sine and cosine functions as in Equations (3.4, 3.5 3.4)

4 SIMULATION STUDY

The main goal of this project was to assess the performance of the various hierarchical clustering algorithms (Single, Average and Complete) using the Fourier basis smoother with and without a weighing function. In our simulations, we had two continuous variables, one functional variable, one directional variable and a categorical variable. With 1000 iterations and cluster number of four, the cluster membership varied from being equal i.e 25 per cluster to different cluster numbers under various conditions.

4.1 Variable Set up In **R**

4.1.1 Categorical

The function sample.int(n, size = n, replace = FALSE, prob=NULL)) in **R** was used to simulate the categorical variables with **n** being the number of continuous variables in each cluster, **size** is the cluster membership, **replace** was set to **TRUE** to sample with replacement and a **prob** for a vector of probability weights. For the four clusters in our study we assigned the probability vectors as follows:

Case 1: **prob**=c(0.8,0.05,0.05,0.05,0.05)

Case 2: **prob**=c(0.05,0.8,0.05,0.05,0.05)

Case 3: **prob**=c(0.05,0.05,0.8,0.05,0.05)

Case 4: **prob**=c(0.08,0.05,0.05,0.8,0.05)

and a probability vector of (0.2,0.2,0.2,0.2,0.2) for an equally likely category.

4.1.2 Functional

The simulation of the functional data followed those of Ferriera and Hitchcock [9] which was also adopted by Hendrickson [18], Chen, Reiss and Tarpey [34], Koomson [25] and Oppong [2]. Ferriera and Hitchcock defined the functional signal groups as follows:

1. Involving periodic tendencies

$$\mu_1(t) = \frac{1}{28}t + e^{-t} + \frac{1}{5}\sin\frac{t}{3} + \frac{1}{2},\ t \in [0, 100]$$
$$\mu_2(t) = \frac{1}{20}t + e^{-t} + \frac{1}{5}\sin\frac{t}{2},\ t \in [0, 100]$$
$$\mu_3(t) = \frac{1}{15}t + e^{-t} + \frac{1}{5}\cos\frac{t}{2} - 1,\ t \in [0, 100]$$
$$\mu_4(t) = \frac{1}{18}t + e^{-t} + \frac{1}{5}\cos\frac{t}{2},\ t \in [0, 100]$$

2. With no periodic tendencies

$$\mu_1(t) = 50 - \frac{t^2}{500} - 7\ln t,\ t \in (0, 100]$$
$$\mu_2(t) = 50 - \frac{t^2}{500} - 5\ln t,\ t \in (0, 100]$$
$$\mu_3(t) = 50 - \frac{t^2}{750} - 7\ln t,\ t \in (0, 100]$$
$$\mu_4(t) = 50 - \frac{t^2}{250} - 4\ln t,\ t \in (0, 100]$$

3. Involving a mixture of periodic and strictly decreasing tendencies

$$\mu_1(t) = -\frac{t}{2} + 2\sin\frac{t}{5},\ t \in (0, 100]$$
$$\mu_2(t) = -\frac{t}{2} + 2\cos\frac{t}{3},\ t \in (0, 100]$$
$$\mu_3(t) = -\frac{t^2}{250} - 4\ln t,\ t \in (0, 100]$$
$$\mu_4(t) = -\frac{t^2}{250} - 2\ln t,\ t \in (0, 100]$$

It should be noted that for this work, we used the signal group involving periodic tendencies for our simulations. These signal functions were chosen so we could get clusters with good representation which are not monotonic. With these functions, we simulated the time vector over the range of 0 to 100 as described by Ferriera and Hitchcock [9] with 0.5 increments. To mimic the natural variations associated with data, we introduced random error terms to our simulated data using a process known as the Ornstein-Uhlenbeck process. The Ornstein-Uhlenbeck is a process that follows a continuous univariate Markov chain evolving over time [13] with zero mean and a defined covariance between error terms over time i and j as

$$\sum = \frac{\sigma^2}{2\beta} e^{(-\beta|t_i - t_j|)} \tag{4.1}$$

where β is the drift variable, which we kept at 0.5 and σ is the variation component which we set at 1.75 and 1 for small and large distances between clusters respectively as used by Hendrickson [18]. To smooth the simulated data after the introduction

of the random error terms, we applied the Fourier basis as described in section (3.3) with the number of basis set to 5. In **R** we used the function **create.fourier.basis (rangeval, nbasis)** where **rangeval** is a 2-length vector that contains the initial and final range values of the functional data being evaluated and **nbasis** is the number of basis. Also, it should be noted that for the functional variables, simulations 1, 2, 3, 5, 6, 7, 10 and 12 contained data that had large distances between the clusters with varying distances between the clusters for the other variables.

4.1.3 Continuous

We simulated the continuous variables from the normal distribution with mean μ and variance σ^2. In similarity to that of Hendrickson (2014) we set the mean at standard deviations for each cluster and values of k being 5,20 and 50 as indicated in Table 4.1.

Table 4.1: Assigned Cluster Mean and Standard Deviation

Category	μ	σ
Cluster 1	5000	100
Cluster 2	5000+$k\sigma$	100
Cluster 3	5000+2$k\sigma$	100
Cluster 4	5000+3$k\sigma$	100

In **R** we used the function abs(rnorm(n,mean,sd)). For this study n is the size of cluster membership.

4.1.4 Directional

To simulate the directional variable the Von Mises distribution [35] defined as

$$\phi(\theta) = (2\pi I_0(\kappa))^{-1} \exp\left(\kappa \cos(\theta - \mu)\right),\ 0 \leq \theta \leq 2\theta,\ 0 \leq \mu < 2\pi,\ \kappa \geq 0 \tag{4.2}$$

was used with $I_0(\kappa)$ as the Bassels function [3] defined as

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp \kappa \cos\theta d\theta \tag{4.3}$$

In **R** we used the function **rvonmises(n, m, κ, rads = TRUE)** where **n** is the number of cluster membership, **m** is the mean which we set at 0 for cluster 1, $0 + k$ for cluster 2, $0 + 2k$ for cluster 3 and $0 + 3k$ for cluster 4, $\kappa = 50$ and **rads** is set to **TRUE** if the mean angle is expressed in radians and **FALSE** otherwise. The values of k were $\frac{1}{2}, 1, \frac{5}{2}$.

4.2 Weight Functions

In this study we will consider the proposed weighting function of Chen, Reiss and Tarpey [34] known as the CV optimal weighting. Other weighing methods include the inverse of the covariance matrix of error terms and the inverse weight function. It is noted that the most common weighting scale used in mixed data clustering has been the inverse variance function which is defined as

$$w(t) = \frac{\frac{1}{\hat{\sigma}^2(t)}}{\int \frac{1}{\hat{\sigma}^2(u)} du} \tag{4.4}$$

where $\hat{\sigma}^2(t)$ is the estimate of the variance of $\theta(t)$. The variance-covariance matrix or otherwise known as the covariance matrix is a measurement to check the correlation between variables. The diagonals of the matrix are made up of the variances whiles the off-diagonal represent the covariance between the variables. For two variables X and Y the covariance denoted as $Cov(X, Y)$ is defined as

$$Cov[X, Y] = E[XY] - E[X]E[Y] \tag{4.5}$$

In matrix notation the Covariance matrix is of the form

$$\sum = \begin{bmatrix} \sigma^2(x_1) & cov[x_1, x_2] & ... & cov[x_1, x_n] \\ cov[x_2, x_1] & \sigma^2(x_2) & ... & cov[x_2, x_n \\ ... & cov[x_3, x_2] & ... & cov[x_3, x_n] \\ cov[x_n, x_1] & ... & ... & \sigma^2(x_n) \end{bmatrix}$$

where $\sum$ is the variance-covariance matrix of nxn dimension.

4.2.1 CV-Optimal Weight

The CV optimal weight function was proposed by Chen, Reiss and Tarpey [34] purposely to smooth and minimize the effect of the coefficient of variation (CV) of

$$||\theta||_w^2 = \sqrt{\int w(t)\theta(t)^2 dt} \tag{4.6}$$

with random function $\theta(t) = b(t)^T\mathbf{z}$ given a $K-$ dimensional vector $\mathbf{z}$ with $b(t) = [b_1(t)...., b_k(t)]^T$ and $[b_1, ..., b_k]$ representing basis functions defined on the interval $[L, U] \subset \mathbf{R}$. Here, they define θ as the difference between the i_{th} and j_{th} of

a set of observed functions $x_1, ..., x_n$. Chen, Reiss and Tarpey [34] argued that the independence assumption underlying the the development of the L_2 metric is unrealistic for most functional data and its applications hence their proposal. They defined the CV-optimal weight as

$$w(t) = [\mathbf{b}_w^{\mathbf{T}}\mathbf{q}]^2 \tag{4.7}$$

where $\mathbf{b}_w(t) = [b_{w1}(t), b_{w2}(t), ..., b_{wk_w}(t)]^{\mathbf{T}}$ is a k_w- dimensional spline basis with associated vector $\mathbf{q}$. In our analysis we applied the CV optimal weight to our functional data and calculated the Rand index for the various clustering algorithms. We also did an analysis without weights so we could compare whether the weight function improves the clustering solutions or not. To apply weights to our simulated functional data we used the **metric.lp(fdata, w=1)** function in **R** where **fdata** is the functional data under study and **w** is the vector of weights. If **w**=1, then the functional data is unweighted.

4.3 Rand Index and Adjusted Rand Index

The Rand index measures the similarity between clustering algorithms and tell the researcher which method is best. For a good clustering algorithm, a high Rand value is expected and vice versa. In this study we shall use the Rand index to test the performance of the single, average and complete linkage algorithms. The Rand index is defined as

$$R = \frac{a+b}{a+b+c+d},\ 0 \leq R \leq 1 \tag{4.8}$$

where:

- a is a pair of subsets placed in the same cluster by clustering method I and clustering method II
- b is a pair of subsets placed in different clusters by clustering method I and clustering method II
- c is a pair of subsets placed in the same cluster with clustering method I but is in a different cluster with clustering method II
- d is a pair of subsets placed in same cluster by clustering method II but in a different cluster with clustering method I.

Another form of similarity measure to check the performance of a clustering algorithm is the adjusted Rand index (ARI). A high value of the adjusted Rand index implies that, there is similarity between the data points in the clustering algorithm and a low value means the data points do not have much similarity or were assigned randomly to form part of the cluster. Unlike the Rand index, the adjusted Rand index can take on negative values. A negative adjusted Rand index value means that there is no random selection or similarity between the data points, rather there is some sort of underlying pattern between them. The adjusted Rand index is defined as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{0.5[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}} \tag{4.9}$$

where n_{ij} is the the number of object in both cluster A_i and B_j, A_i is the i_{th} cluster in clustering method II and B_j is the j_{th} cluster in clustering method II and $\sum_{ij} = N$.

4.4 Monte Carlo Standard Error

The Monte Carlo standard error (MCSE) is a measure of accuracy used when several simulations are done in a study. Ideally, the probability of varying outcomes is indeterminable due to the interference of random variables. So, the MCSE repeatedly collect errors in the simulation process after which the results are averaged then an estimate is made. This measure further tells how large our estimation noise is. We will use the MCSE to measure the variability of the Rand index across all simulation studies. We observed that the MCSE values are very small so we rely on the mean Rand and mean Adjusted Rand for our comparison. However the MCSE values will be presented in the comparison tables.

4.5 Simulation Results

The means of the unweighted Rand and adjacent Rand indexes are presented in Tables (4.2-4.5) along with their respective Monte Carlo standard errors. The performance of the extended Gower across the average and complete linkage methods were generally good as most of the Rand and adjusted Rand values were high but that was not the case for the single linkage method. With regards to ranks in performance, the single linkage performed worse across every simulation stage compared to the other two algorithms. We observed that in some instances the average linkage performs better than the complete linkage and vice versa. However, the approximate differences between them is not significant enough to clearly state which algorithm stands out. We also observed that where equal cluster size is allocated, the average linkage performs slightly better than the complete linkage even though that was not for all

cases. This suggests that to cluster using equal cluster sizes the average linkage is the preferred choice. In almost every simulation setting with a cluster size allocation of (10, 20, 30 ,40) and (33, 33, 33, 1), the complete linkage method performs better. When we applied the weight function we a saw a general improvement in the Rand values with the single linkage method being the most improved. However, it still performed poorly compared to the other two methods. Contrary to the observed improvement in Rand values after the weights were applied, the unweighted Rand values for simulation 14 (a,b) were slightly higher than the Rand values for the weighted simulated values. The variations within clusters were statistically good because the MCSE values were low (< 0.02). Results for the weighted Rand and adjacent Rand index is presented in Tables (A.1-A.4) in the appendix.

Table 4.2: Rand Comparison for Simulation 1a-4b: Unweighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
1a	Equal size	Single	0.5192	0.2133	0.0051
		Average	0.9706	0.9202	0.0007
		Complete	0.9566	0.8920	0.0148
1b	(33,33,33,1)	Single	0.4981	0.1822	0.0054
		Average	0.9633	0.9152	0.0009
		Complete	0.9640	0.9160	0.0010
1c	(10,20,30,40)	Single	0.5407	0.2218	0.0058
		Average	0.9714	0.9311	0.0008
		Complete	0.9634	0.9132	0.0013
2a	Equal sizes	Single	0.5207	0.2273	0.0052
		Average	0.8662	0.7079	0.0014
		Complete	0.8661	0.6640	0.0016
2b	(33,33,33,1)	Single	0.5047	0.1832	0.0051
		Average	0.8694	0.7064	0.0012
		Complete	0.8768	0.7159	0.0019
2c	(10,20,30,40)	Single	0.5482	0.2211	0.0053
		Average	0.8727	0.6971	0.0016
		Complete	0.8727	0.6971	0.0017
3a	Equal Sizes	Single	0.4128	0.0617	0.0013
		Average	0.9184	0.7872	0.0018
		Complete	0.8886	0.7219	0.0024
3b	(33,33,33,1)	Single	0.4537	0.0737	0.0012
		Average	0.9165	0.8022	0.0018
		Complete	0.9164	0.8020	0.0018
3c	(10,20,30,40)	Single	0.4415	0.0217	0.0018
		Average	0.9322	0.7824	0.0056
		Complete	0.9120	0.7834	0.0024
4a	Equal sizes	Single	0.5018	0.2110	0.0005
		Average	0.9205	0.2110	0.0013
		Complete	0.9025	0.7563	0.0019
4b	(33,33,33,1)	Single	0.5020	0.1825	0.0052
		Average	0.9008	0.8211	0.0017
		Complete	0.9013	0.8157	0.0014

Table 4.3: Rand Comparison for Simulation 4c-8a: Unweighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
4c	(10,20,30,40)	Single	0.5269	0.2015	0.0055
		Average	0.8945	0.7978	0.0018
		Complete	0.9100	0.7951	0.0015
5a	Equal size	Single	0.5093	0.2097	0.0046
		Average	0.9487	0.8947	0.0012
		Complete	0.9547	0.8988	0.0018
5b	(33,33,33,1)	Single	0.4623	0.2137	0.0056
		Average	0.9682	0.9146	0.0018
		Complete	0.9532	0.9226	0.0011
5c	(10,20,30,40)	Single	0.5322	0.2256	0.0051
		Average	0.9624	0.9099	0.0113
		Complete	0.9644	0.9100	0.0113
6a	Equal sizes	Single	0.5476	0.2781	0.0051
		Average	0.7496	0.3781	0.0001
		Complete	0.7496	0.3780	0.0001
6b	(33,33,33,1)	Single	0.4629	0.2244	0.0034
		Average	0.7130	0.5886	0.0014
		Complete	0.7131	0.3248	0.0014
6c	(10,20,30,40)	Single	0.4174	0.3179	0.0012
		Average	0.7196	0.3181	0.0012
		Complete	0.7776	0.3183	0.0012
7a	Equal Sizes	Single	0.5481	0.2111	0.0035
		Average	0.8484	0.6115	0.0015
		Complete	0.8483	0.6113	0.0015
7b	(33,33,33,1)	Single	0.4441	0.2397	0.0043
		Average	0.8443	0.6402	0.0019
		Complete	0.8446	0.6404	0.0019
7c	(10,20,30,40)	Single	0.4456	0.3376	0.0007
		Average	0.8427	0.6308	0.0015
		Complete	0.8457	0.6308	0.0015
8a	Equal sizes	Single	0.4986	0.2546	0.0013
		Average	0.8898	0.7254	0.0024
		Complete	0.8899	0.7254	0.0024

Table 4.4: Rand Comparison for Simulation 8b-11c: Unweighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
8b	(33,33,33,1)	Single	0.4334	0.0651	0.0034
		Average	0.9137	0.7959	0.0019
		Complete	0.9136	0.7958	0.0019
8c	(10,20,30,40)	Single	0.4449	0.0785	0.0034
		Average	0.9102	0.7962	0.0023
		Complete	0.9133	0.7964	0.0023
9a	Equal size	Single	0.4124	0.0610	0.0012
		Average	0.8879	0.7065	0.0014
		Complete	0.8648	0.6609	0.0016
9b	(33,33,33,1)	Single	0.5019	0.1782	0.0051
		Average	0.8061	0.1782	0.0019
		Complete	0.8730	0.0051	0.0019
9c	(10,20,30,40)	Single	0.5587	0.2329	0.0052
		Average	0.8818	0.6464	0.0015
		Complete	0.8728	0.6963	0.0015
10a	Equal sizes	Single	0.4146	0.0487	0.0016
		Average	0.7709	0.4781	0.0011
		Complete	0.7923	0.4781	0.0011
10b	(33,33,33,1)	Single	0.4505	0.0607	0.0014
		Average	0.7051	0.5218	0.0020
		Complete	0.7151	0.5219	0.0020
10c	(10,20,30,40)	Single	0.4589	0.0206	0.0022
		Average	0.7747	0.4502	0.0018
		Complete	0.7746	0.4500	0.0018
11a	Equal Sizes	Single	0.5127	0.2022	0.0053
		Average	0.9069	0.7645	0.0019
		Complete	0.9069	0.7646	0.0019
11b	(33,33,33,1)	Single	0.4948	0.1729	0.0051
		Average	0.8111	0.8186	0.0014
		Complete	0.9101	0.9216	0.0015
11c	(10,20,30,40)	Single	0.5254	0.5254	0.0055
		Average	0.9131	0.7918	0.0015
		Complete	0.9131	0.7982	0.0015

Table 4.5: Rand Comparison for Simulation 12a-15b: Unweighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
12a	Equal sizes	Single	0.5552	0.0096	0.0001
		Average	0.7128	0.2589	0.0001
		Complete	0.7129	0.2592	0.0009
12b	(33,33,33,1)	Single	0.5472	0.0084	0.0014
		Average	0.5821	0.1086	0.0015
		Complete	0.6256	0.1086	0.0015
12c	(10,20,30,40)	Single	0.5205	0.0007	0.0012
		Average	0.6635	0.1881	0.0012
		Complete	0.6636	0.1882	0.0012
13a	Equal size	Single	0.5229	0.0190	0.0024
		Average	0.6947	0.0211	0.0016
		Complete	0.7377	0.3401	0.0006
13b	(33,33,33,1)	Single	0.4930	0.0279	0.0016
		Average	0.5919	0.0316	0.0008
		Complete	0.7127	0.3237	0.0015
13c	(10,20,30,40)	Single	0.5421	0.0079	0.0018
		Average	0.6696	0.3191	0.0012
		Complete	0.7185	0.2053	0.0010
14a	Equal sizes	Single	0.5864	0.2597	0.0047
		Average	0.8491	0.6141	0.0015
		Complete	0.8492	0.6143	0.0015
14b	(33,33,33,1)	Single	0.5565	0.2318	0.0052
		Average	0.8443	0.6403	0.0019
		Complete	0.8835	0.6404	0.0019
14c	(10,20,30,40)	Single	0.5426	0.2248	0.0037
		Average	0.8428	0.6251	0.0015
		Complete	0.8428	0.6252	0.0015
15a	Equal sizes	Single	0.4776	0.4562	0.0018
		Average	0.7770	0.45658	0.0018
		Complete	0.7771	0.4568	0.0018
15b	(33,33,33,1)	Single	0.4608	0.0493	0.0027
		Average	0.7000	0.5304	0.0020
		Complete	0.7991	0.5307	0.0021

Table 4.6: Rand Comparison for Simulation 15c: Unweighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
15c	(10,20,30,40)	Single	0.4676	0.0231	0.0019
		Average	0.7034	0.3901	0.0015
		Complete	0.7732	0.4469	0.0018

5 DISCUSSION / FUTURE RESEARCH

This thesis work was carried out to compare hierarchical clustering methods (single, average and complete) and assess their performance with functional data with mixed attributes using the Fourier smoothing technique along with a weighted or unweighted function.

We start by introducing clustering methods and the improvements made over the years to include mixed data types into the clustering algorithm for better representation and understanding. We did a simulation with 1000 iterations and added noise using the Ornstein-Uhlenbeck process then proceeded to smooth the data with the Fourier basis function. We considered the CV optimal weight function [34] which is designed to minimize the coefficient of variation between data with functional attributes. We also employed the extension of the Gower coefficient to accommodate functional, continuous, categorical and directional variables in our simulated datasets.

To assess how the algorithms perform under different structure, we used different clustering sizes. First we considered the case where each sample cluster is of the same size of 25 and moved on to when a cluster has a maximum size of 33 and minimum of 1 etc. In comparing the performance of the various clustering methods we calculated the Rand index and adjusted Rand index of our simulated data.

In general we saw an improvement in the weighted approach as compared to the standard unweighted approach. However, this was not so for all cases for simulation 14 (a,b). We observed that the performance of the extended Gower coefficient over all the setting produced reasonable good results.

For future research and development of clustering mixed data, a comparative anal-

ysis of how smoothed functional data compares to when functional data is analyzed without smoothing can be done to ascertain how the results compare. For this work, we focused on removing some of the noise in the data before clustering and would be interesting to see how the results compare to those when we do not remove the noise. We suggest that attention be geared towards correlated functional variables of mixed structure. Also, a statistical method for simulating mixed data with cluster structures that can accommodate more than two clusters should be explored.

BIBLIOGRAPHY

[1] Hanns Ackermann. A note on circular nonparametrical classification. *Biometrical journal*, 39(5):577–587, 1997.

[2] Oppong Augustine. Clustering mixed data: An extension of the gower coefficient with weighted l 2 distance, 2018.

[3] DJ Best and Nicholas I Fisher. Efficient simulation of the von mises distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(2):152–157, 1979.

[4] Huaihou Chen, Philip T Reiss, and Thaddeus Tarpey. Optimally weighted l2 distance for functional data. *Biometrics*, 70(3):516–525, 2014.

[5] Abhijit Dasgupta and Adrian E Raftery. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American statistical Association*, 93(441):294–302, 1998.

[6] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[7] Marcello D'Orazio. Distances with mixed type variables some modified gower's coefficients. *arXiv preprint arXiv:2101.02481*, 2021.

[8] Brian S Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & probability letters*, 6(5):305–309, 1988.

[9] Laura Ferreira and David B. Hitchcock. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics - Simulation and Computation*, 38:1925 – 1949, 2009.

[10] Nicholas I Fisher. *Statistical analysis of circular data*. cambridge university press, 1995.

[11] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[12] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association.*, 97(458), 2002.

[13] Daniel T Gillespie. Exact numerical simulation of the ornstein-uhlenbeck process and its integral. *Physical review E*, 54(2):2084, 1996.

[14] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27(4):857–871, 1971.

[15] Christian Hennig, Marina Meila, Fionn Murtagh, and Rocci Roberto. *Handbook of cluster analysis*. CRC Press, 2016.

[16] Jih-Jeng Huang, Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. Marketing segmentation using support vector clustering. *Expert systems with applications*, 32(2):313–317, 2007.

[17] Anil K Jain and Richard C Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[18] Hendrickson JeanMarie. *Methods for clustering mixed data*. PhD thesis, University of South Carolina, 2014.

[19] Jong-Min Chae Kim, San Seong, and Wan Youn Yang. Cluster analysis with balancing weight on mixed-type data. *The Korean Communications in Statistics*, 13(3), 2006.

[20] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[21] Kaufman Leonard and Rousseeuw Peter J. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

[22] G McLachlan and K Thriyambakam. The em algorithm and extensions new york wiley. 1997.

[23] Melnykov, Volodymyr, and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Publications. 67.*, 2010.

[24] Van de Velden Michel, Iodice D'Enza Alfonso, and Markos Angelos. *Distance-based clustering of mixed data*. Wires Computational Statistics, 2018.

[25] Koomson Obed. Performance assessment of the extended gower coefficient on mixed data with varying types of functional data, 2018.

[26] Jaccard Paul. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 44:223–70, 01 1908.

[27] Amanda Prokasky, Kathleen Rudasill, Victoria J. Molfese, Samuel Putnam, Maria Gartstein, and Mary Rothbart. Identifying child temperament types using cluster analysis in three samples. *Journal of Research in Personality*, 67:190–201, 2017. Personality in Childhood.

[28] Essomba Rene. Smoothing Techniques using basis functions: Fourier Basis. `https://datascienceplus.com/smoothing-techniques-using-basis-functions-fourier-basis/`, 2015.

[29] Lior Rokach and Oded Maimon. *Clustering Methods*, pages 321–352. 01 2005.

[30] I Rytsarev, Alexander Kupriyanov, D Kirsh, and K Liseckiy. Clustering of social media content with the use of bigdata technology. *Journal of Physics: Conference Series*, 1096:012085, 09 2018.

[31] Everitt Brian S, Landau Sabine, Leese Morven, and Stahl Daniel. Cluster analysis 5th ed, 2011.

[32] Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

[33] R.R. Sokal, C.D. Michener, and University of Kansas. *A Statistical Method for Evaluating Systematic Relationships*. University of Kansas science bulletin. University of Kansas, 1958.

[34] Thaddeus Tarpey. Linear Transformations and the k-Means Clustering Algorithm: Applications to Clustering Curves. *The American Statistician*, 61:34–40, February 2007.

[35] Richard von Mises. About the "integer" of atomic weights and related questions. *physical. Z.*, 19:490–500, 1981.

[36] Joseph Zubin. A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, 33(4):508, 1938.

APPENDICES

A Weighted Rand Index Comparison

Table A.1: Rand Comparison for Simulation 1a-4b: Weighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
1a	Equal size	Single	0.5366	0.2309	0.0051
		Average	0.9727	0.9258	0.0006
		Complete	0.9592	0.8964	0.0016
1b	(33,33,33,33,1)	Single	0.5081	0.1928	0.0054
		Average	0.9627	0.9138	0.0009
		Complete	0.9628	0.9137	0.0011
1c	(10,20,30,40)	Single	0.5444	0.2288	0.0057
		Average	0.9728	0.9344	0.0008
		Complete	0.9637	0.9128	0.0013
2a	Equal size	Single	0.5475	0.2140	0.0052
		Average	0.8888	0.7075	0.0014
		Complete	0.8685	0.6699	0.0016
2b	(33,33,33,1)	Single	0.5112	0.1902	0.0051
		Average	0.8694	0.8730	0.0014
		Complete	0.8718	0.8818	0.0019
2c	(10,20,30,40)	Single	0.5495	0.2235	0.0053
		Average	0.8751	0.6994	0.0012
		Complete	0.8770	0.7070	0.0017
3a	Equal size	Single	0.4148	0.0631	0.0012
		Average	0.9188	0.7885	0.0018
		Complete	0.8877	0.7200	0.0024
3b	(33,33,33,1)	Single	0.4559	0.0758	0.0012
		Average	0.9213	0.8158	0.0017
		Complete	0.9138	0.9198	0.0018
3c	(10,20,30,40)	Single	0.4489	0.0150	0.0017
		Average	0.9476	0.8739	0.0016
		Complete	0.9166	0.7947	0.0023
4a	Equal size	Single	0.5153	0.2063	0.0053
		Average	0.9233	0.7880	0.0011
		Complete	0.9033	0.7534	0.0018
4b	(33,33,33,1)	Single	0.5020	0.1662	0.0050
		Average	0.9100	0.7920	0.0012
		Complete	0.9180	0.8104	0.0015

Table A.2: Rand Comparison for Simulation 4c-8a: Weighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
4c	(10,20,30,40)	Single	0.5314	0.2062	0.0054
		Average	0.9116	0.7869	0.0012
		Complete	0.9123	0.7900	0.0015
5a	Equal size	Single	0.5351	0.2288	0.0051
		Average	0.9726	0.9257	0.0068
		Complete	0.9578	0.8933	0.0016
5b	(33,33,33,1)	Single	0.5003	0.1840	0.0057
		Average	0.9698	0.9164	0.0009
		Complete	0.9623	0.9125	0.0010
5c	(10,20,30,40)	Single	0.5490	0.2312	0.0058
		Average	0.9723	0.9331	0.0008
		Complete	0.9649	0.9159	0.0012
6a	Equal size	Single	0.1124	0.0553	0.0014
		Average	0.7267	0.3115	0.0007
		Complete	0.7485	0.3775	0.0005
6b	(33,33,33,33,1)	Single	0.4716	0.0394	0.0015
		Average	0.5886	0.7130	0.0001
		Complete	0.7167	0.3326	0.0013
6c	(10,20,30,40)	Single	0.5189	0.7200	0.0021
		Average	0.6664	0.1983	0.0010
		Complete	0.7189	0.3220	0.0013
7a	Equal size	Single	0.5776	0.2507	0.0048
		Average	0.8426	0.6012	0.0017
		Complete	0.8480	0.6115	0.0014
7b	(33,33,33,1)	Single	0.5404	0.2144	0.0053
		Average	0.8340	0.6207	0.0019
		Complete	0.8441	0.6387	0.0018
7c	(10,20,30,40)	Single	0.6210	0.2995	0.0047
		Average	0.8427	0.6229	0.0014
		Complete	0.8459	0.6320	0.0016
8a	Equal size	Single	0.4117	0.6090	0.0018
		Average	0.9203	0.7913	0.0017
		Complete	0.8858	0.7150	0.0021

Table A.3: Rand Comparison for Simulation 8b-11c: Weighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
8b	(33,33,33,1)	Single	0.5533	0.7340	0.0012
		Average	0.9232	0.8196	0.0017
		Complete	0.9128	0.7915	0.0019
8c	(10,20,30,40)	Single	0.5474	0.0214	0.0018
		Average	0.9447	0.8674	0.0017
		Complete	0.9161	0.7904	0.0023
9a	Equal size	Single	0.5342	0.2186	0.0051
		Average	0.8866	0.7027	0.0014
		Complete	0.8688	0.6699	0.0016
9b	(33,33,33,1)	Single	0.5058	0.1832	0.0051
		Average	0.8691	0.6970	0.0014
		Complete	0.8730	0.8730	0.0018
9c	(10,20,30,40)	Single	0.5626	0.2276	0.0053
		Average	0.8807	0.7128	0.0012
		Complete	0.8710	0.6993	0.0015
10a	Equal size	Single	0.4177	0.0502	0.0017
		Average	0.7749	0.4277	0.0010
		Complete	0.7997	0.4749	0.0011
10b	(33,33,33,1)	Single	0.4569	0.0597	0.0014
		Average	0.7460	0.4153	0.0010
		Complete	0.7981	0.5281	0.0021
10c	(10,20,30,40)	Single	0.4619	0.0212	0.0022
		Average	0.7747	0.3628	0.0015
		Complete	0.7758	0.4529	0.0018
11a	Equal size	Single	0.5198	0.2084	0.0053
		Average	0.9174	0.7801	0.0012
		Complete	0.9068	0.7644	0.0018
11b	(33,33,33,33,1)	Single	0.4948	0.1728	0.0051
		Average	0.9080	0.7875	0.0012
		Complete	0.9214	0.8185	0.0016
11c	(10,20,30,40)	Single	0.5264	0.5264	0.0053
		Average	0.9057	0.9057	0.0012
		Complete	0.9204	0.8160	0.0015

Table A.4: Rand Comparison for Simulation 12a-15b: Weighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
12a	Equal size	Single	0.5534	0.0092	0.0023
		Average	0.6355	0.0555	0.0010
		Complete	0.7148	0.2652	0.0010
12b	(33,33,33,1)	Single	0.5485	0.0080	0.0013
		Average	0.5821	0.6256	0.0004
		Complete	0.6291	0.1181	0.0016
12c	(10,20,30,40)	Single	0.5698	0.0021	0.0013
		Average	0.6653	0.0537	0.0008
		Complete	0.6690	0.1884	0.0012
13a	Equal size	Single	0.5482	0.2237	0.0054
		Average	0.8377	0.6289	0.0019
		Complete	0.6935	0.2094	0.0011
13b	(33,33,33,1)	Single	0.4974	0.0284	0.0054
		Average	0.5897	0.0277	0.0087
		Complete	0.7121	0.8576	0.0014
13c	(10,20,30,40)	Single	0.5428	0.0060	0.0018
		Average	0.6680	0.2026	0.0011
		Complete	0.7193	0.3219	0.0012
14a	Equal size	Single	0.5678	0.2422	0.0051
		Average	0.8474	0.6111	0.0017
		Complete	0.8479	0.6113	0.0015
14b	(33,33,33,1)	Single	0.5442	0.2183	0.0053
		Average	0.8359	0.6246	0.0018
		Complete	0.8415	0.6351	0.0020
14c	(10,20,30,40)	Single	0.6194	0.2972	0.8428
		Average	0.8446	0.6293	0.0015
		Complete	0.8479	0.6113	0.0015
15a	Equal size	Single	0.4467	0.4767	0.0022
		Average	0.7337	0.7337	0.0014
		Complete	0.7780	0.7780	0.0018
15b	(33,33,33,1)	Single	0.4608	0.0558	0.0015
		Average	0.7000	0.4867	0.0032
		Complete	0.7770	0.4469	0.0021

Table A.5: Rand Comparison for Simulation 13a-15c: Weighted

Simulation	Cluster Allocation	Method	Mean Rand	Mean Adjusted Rand	MCSE
15c	(10,20,30,40)	Single	0.4728	0.0192	0.0021
		Average	0.7347	0.3609	0.0014
		Complete	0.7742	0.4479	0.0018

VITA

ISHMAEL AMARTEY

Education:	M.S. Mathematical Science, East Tennessee State University, Johnson City, Tennessee 2021 B.S. Actuarial Science, University for Development Studies, Tamale, Ghana 2014
Professional Experience:	Graduate Teaching Assistant, East Tennessee State University, Johnson City, Tennessee, 2020–2021 Sales Associate, Prudential Life Insurance Ghana, Accra, Ghana 2019–2020 Director of Education, Sanitation and Environmental Development Community, Accra, Ghana 2015–2019 Public Relations Officer, National Health Insurance Authority, Ashiedu-Keteke District Accra, Ghana 2014–2015
Professional Development:	Statistical and Mathematical Software: R, SPSS Microsoft Office Suite: Word, Excel, PowerPoint, Outlook