View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by East Tennessee State University

SCHOOL of GRADUATE STUDIES
EAST TENNESSEE STATE UNIVERSITY

East Tennessee State University
Digital Commons @ East Tennessee State University

Electronic Theses and Dissertations

Student Works

12-2018

Performance Assessment of The Extended Gower Coefficient on Mixed Data with Varying Types of Functional Data.

Obed Koomson
East Tennessee State University

Follow this and additional works at: https://dc.etsu.edu/etd

Part of the Applied Mathematics Commons

Recommended Citation

Koomson, Obed, "Performance Assessment of The Extended Gower Coefficient on Mixed Data with Varying Types of Functional Data." (2018). *Electronic Theses and Dissertations.* Paper 3512. https://dc.etsu.edu/etd/3512

This Thesis - Open Access is brought to you for free and open access by the Student Works at Digital Commons @ East Tennessee State University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ East Tennessee State University. For more information, please contact digilib@etsu.edu.

Performance Assessment of The Extended Gower Coefficient on Mixed Data with Varying Types of Functional Data

---

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

Obed Koomson

December 2018

JeanMarie Hendrickson, Ph.D., Chair

Robert Price, Ph.D.,

Nicole Lewis , Ph.D.

Keywords: Hierarchical Clustering, Mixed data, Strictly-decreasing signal, Periodic signal, Extended Gower coefficient.

ABSTRACT

Performance Assessment of The Extended Gower Coefficient on Mixed Data with Varying Types of Functional Data

by

Obed Koomson

Clustering is a widely used technique in data mining applications to source, manage, analyze and extract vital information from large amounts of data. Most clustering procedures are limited in their performance when it comes to data with mixed attributes. In recent times, mixed data have evolved to include directional and functional data. In this study, we will give an introduction to clustering with an eye towards the application of the extended Gower coefficient by Hendrickson (2014). We will conduct a simulation study to assess the performance of this coefficient on mixed data whose functional component has strictly-decreasing signal curves and also those whose functional component has a mixture of strictly-decreasing signal curves and periodic tendencies. We will assess how four different hierarchical clustering algorithms perform on mixed data simulated under varying conditions with and without weights. The comparison of the various clustering solutions will be done using the Rand Index.

Copyright by Obed Koomson, 2018

All Rights Reserved.

ACKNOWLEDGMENTS

I would like to thank the Almighty God. I would also like to thank the members of my advisory committee. Thank you Dr. Hendrickson for the invaluable contributions to this study through every stage. I am very grateful to you for accepting to work with me and for supervising this work. Thank you Dr. Price for accepting to be on the committee and making invaluable suggestions and contributions. To Dr. Lewis, God bless you for your critical inputs and for keeping me on the lookout for all the deadlines. I want to show appreciation to Dr. Robert Bob Gardner, and the entire ETSU math faculty for their excellent teaching and their positive impact. I can't conclude without saying how grateful I am to my family and friends. God richly bless you all.

TABLE OF CONTENTS

APPENDIX: RAND INDICES AND MONTE CARLO STANDARD ERRORS(MCSE) 53

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

The 21st century has witnessed an ever-increasing need for the use of data; from sourcing and managing data, to analyzing and extracting vital information from large amounts of data. Large multivariate data have evolved to be useful in understanding and describing most real life situations. It is however difficult to understand the constituent of these data, and how they relate to the events around us. For this reason, methods of summarizing and extracting relevant information from large datasets are necessary[1]. Over the years, clustering techniques have proved to be very useful in recognizing and discussing the different types of events, objects and people we encounter.

Cluster analysis has become the generic term used to describe all the activities involved in grouping objects (in a data set) into a set of classes (natural groupings) such that two or more objects that belong to the same class are as similar as possible. It should be noted that the clusters (classes) are dissimilar to each other. Most real data situations have various factors of interest which are measured on different scales. As a result, the traditional methods of classifying data types (quantitative and categorical) are no longer the only ways of looking at data. In fields such as linguistics, psychiatry, geology, pattern recognition, taxonomy, soil science, etc., mixed data appears so much that much attention has been given to methods of clustering them by researchers of late. For example, in soil science, the variables describing a soil horizon could include pH (numeric variable), hue (ordinal variable), and the presence of earthworms (binary variable), etc. [1].

According to Hendrickson(2014) [2], most real data contain different types of variables or attributes. For this reason, an important aspect or area of cluster analysis deals with clustering mixed data. Mixed data is a term used when we have data consisting of several variable types, for instance; continuous, categorical, functional, directional, etc. There are two traditional methods of clustering mixed data. The first is to convert all other variables

into the form of the variable with a lot of observations, and then implement a clustering method on the new dataset. The other method involves carrying out separate analysis, with each analysis involving variables of a single type only. These two methods have some serious setbacks which will be discussed in chapter 2 of this work. There is a third method of clustering mixed data which involves combining the variables into one proximity matrix as described by Kaufman and Rousseeuw (1990) [3]. Most of the activities in cluster analysis focus on finding the dissimilarity between two objects, but this could be looked at differently in a way where our interest will be on finding the similarity between objects. In 1971, Gower [4] came up with a coefficient for measuring this kind of similarity between two objects when we have mixed data (Hendrickson 2014) [2]. Gower's work was basically on a measure of similarity that can be used for a mixture of continuous, categorical and nominal variables. Hendrickson (2014) [2] created an extension of the Gower coefficient to include other types of variables. She included a measure of similarity for directional and functional attributes. Her work however considered only functional data with periodic tendencies. In this work, we will look at the performance of the extended Gower coefficient(Hendrickson 2014) [2] when we have other types of functional data, specifically strictly decreasing functions and a mixture of periodic and strictly decreasing functions. We will also look at how four hierarchical clustering algorithms perform on mixed data, and suggest the best hierarchical clustering algorithm for mixed data with functional and directional variables.

2 LITERATURE REVIEW

2.1 Overview of Cluster Analysis

Cluster analysis is a general term for a group of multivariate techniques whose primary purpose is to find groups in data. That is by cluster analysis, we apply techniques to discover or uncover groupings in data from the characteristics they possess such that objects within one cluster are very similar to others in the cluster with respect to some standard metric or criterion. The results of such a process should exhibit internal homogeneity within clusters and heterogeneity between clusters. In general, clustering procedures are classified as either Hierarchical or Nonhierarchical. In Hierarchical clustering, partitioning is not done at a single step but it is done in a series of partitions, which mostly run from one cluster containing all individuals to **n** clusters each containing a single individual, where **n** is the number of objects in our dataset [6]. Hierarchical methods can further be categorized as agglomerative or divisive. Nonhierarchical methods include the partitioning approach, density estimation and mixtures of distributions. Figure 2.1 shows a breakdown of the various clustering algorithms:

The number of ways of partitioning a set of n items into g clusters is given by:

$$N(\mathbf{n},\mathbf{g}) = \frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^{n}$$

For most hierarchical methods, the number of possible clusters are so large even for some small values of **n**. For this reason, the various approaches to clustering permit us to search for a reasonable solution without having to look at all possible arrangements [26]. Agglomerative methods involve a sequential process in which an observation or a cluster is merged into another at each step. That is, the two closest clusters are merged into a new single cluster. Closeness is measured by means of distances, which is often termed as dissimilarities between objects. The most common measure of distance is the Euclidean distance. We define the

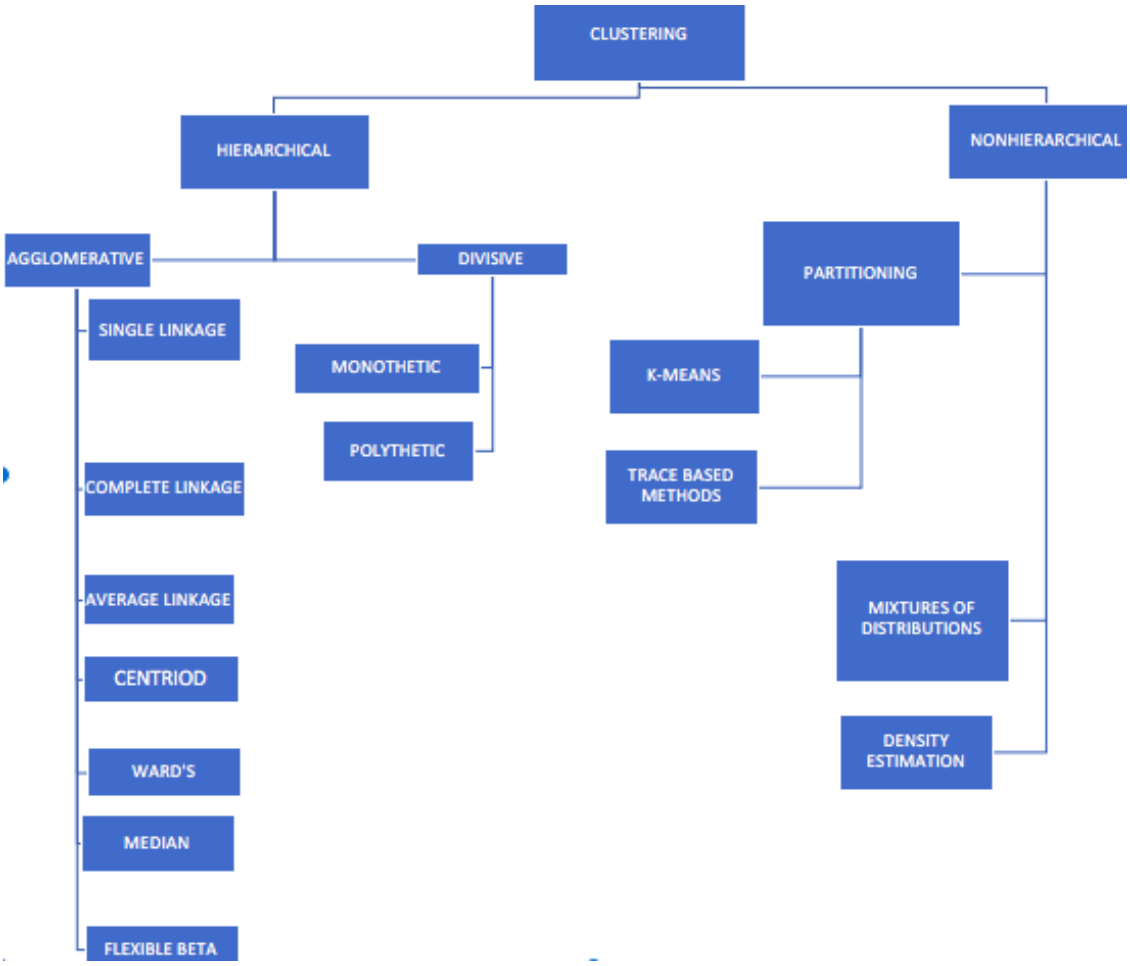

Figure 2.1: Breakdown of the various clustering Algorithms

Euclidean distance between two p-dimensional observations $\mathbf{x}$ and $\mathbf{y}$ as

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

Another common measure of distance used in cluster analysis is the Minkowski distance metric. With this, the distance between two p-dimensional observation $\mathbf{x}$ and $\mathbf{y}$, is given by

$$d(\mathbf{x},\mathbf{y}) = \sum [|x_i - y_i|^p]^{\frac{1}{p}},$$

and another method, the Canberra metric defines the distance between $\mathbf{x}$ and $\mathbf{y}$ as

$$d(\mathbf{x},\mathbf{y}) \sum \frac{|x_i - y_i|}{x_i + y_1}.$$

Different approaches for measuring distances between clusters give rise to the different hierarchical methods. Under agglomerative hierarchical clustering, we have the single linkage method by Sneath (1957) [28], which defines the distance between two clusters A and B as

$$d(A, B) = \min_{i \in A, j \in B} d_{ij}.$$

Another agglomerative method is the complete linkage method which was defined by Sorensen (1948) [19]. For two clusters A and B, Sorensen defined the distance between A and B as

$$d(A, B) = \max_{i \in A, j \in B} d_{ij}.$$

The average linkage method defines the distance between clusters A and B as

$$d(A, B) = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij},$$

where n_A is the number of objects in cluster A, n_B is the number of objects in cluster B, and d_{ij} is the distance between objects.

In the centroid method, the distance between two clusters A and B is defined as the Euclidean distance between the mean vectors of the two clusters and is given by

$$d(A, B) = d(\overline{y}_A, \overline{y}_B),$$

where $\overline{y}_A$ and $\overline{y}_B$ are the mean vectors for the observation vectors in A and the observation vectors in B, respectively.

The median method is used to avoid weighting the mean vectors by cluster size when two clusters A and B are merged, with one containing larger observations than the other. This is a build up from the centroid method and it helps to prevent the situation where the centroid of the combined groups will be closer to the mean of the cluster with the larger amount of observation. It is given as

$$m_{AB} = \frac{1}{2}(\overline{y}_A + \overline{y}_B).$$

Another agglomerative method, the Ward's method, uses the within-cluster squared distances and the between-cluster squared distances [18]. If AB is the cluster obtained by combining clusters A and B, the sum of the within-cluster distances are

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \overline{y}_A)'(y_i - \overline{y}_A),$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \overline{y}_B)'(y_i - \overline{y}_B),$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB})'(y_i - \overline{y}_{AB}),$$

where $\overline{y}_{AB} = (n_A\overline{y}_A + n_B\overline{y}_B)/(n_A + n_B)$, and n_A, n_B, and $n_{AB} = n_A + n_B$ are the numbers of points in A, B, and AB respectively.

All the agglomerative techniques discussed above are very useful techniques. For instance, when clustering functional data, the Ward's method has been identified as the method with the highest accuracy. An example is the work of Ferreira and Hitchcook [7]. We have different types of agglomerative hierarchical clustering algorithms, but all seem to follow a common pattern and they all seem to have one major disadvantage. This has to do with our inability to undo a merger after a grouping is done. For this reason, the merging has to be done carefully so that we don't end up with a low cluster quality.

Under divisive methods, we begin with one cluster which contains the **n** items or objects in our dataset. With this method, we perform a split at each step by dividing the data into two subgroups [26], and at the end of the entire process, we always have **n** clusters. It should be noted that there will always be one item in the clusters we end up with. In general, we have two types of divisive hierarchical clustering. The first method is the monothetic method. With this divisive method, the division of the groups into two subgroups is based on a single variable. The other one is the polythetic method, in which we use all the variables at each stage of splitting. Also with this method, we can work with a proximity matrix.

MacNaughton-Smith et al. (1964) used a method that avoids considering all possible splits [16] . This is however a potential problem of a polythetic method. In their method, they determined the object with the maximum distance from the others within a group. This object so determined, is used as a seed for the splinter group. For the remaining objects, they determined if each of them could be added to the splinter group by means of using the minimum distance between a particular object under consideration and the splinter group. They repeated this with the next cluster for splitting being chosen as the one which is largest in diameter (defined by the largest dissimilarity between any two objects) [6]. One advantage of the divisive method is that, it tends to reveal the data structure from the start of the process. [3].

Under nonhierarchical clustering techniques, the goal is to group items into a collection of **K** clusters. The number of clusters, K, is usually specified in advance or determined as part of the procedure [13]. The most popular nonhierarchical approach to clustering is the K-means method. Under this, we seek K groups at the end of the process by minimizing some criterion. The most widely used criterion is the within-group sum of squares over all variables [6]. K-means was suggested by MacQueen(1967) to describe the process in which he assigned each of the objects to clusters in such a way that, an item gains entry into a cluster if its distance from the centroid of that cluster is shorter than the distance of the item from any other cluster [20]. Some common problems associated with K-means clustering methods include among others: its sensitivity to outliers; specifying the number of groups in advance; and finally it is only applicable in situations where we can compute the cluster means.

Model-based clustering (Finite mixture densities) offers an alternative approach to clustering where we assign a formal statistical model to the population we sampled our data from. With this approach, we assume the existence of sub-populations, which are the groups or clusters. Each of these assumed sub-populations has its own multivariate probability den-

sity function. Because of the existence of several density functions, we often use the term finite mixture densities [6]. The advantage of this method is that it does not require the user to estimate the number of clusters. It also provides some sort of framework for inference. In general, finite mixture densities have a family of probability density functions of the form:

$$f(x; p, \theta) = \sum_{j=1}^{c} p_j g_j(x; \theta_j)$$

where x is a p-dimensional random variable, $\mathbf{p}' = [\mathbf{p_1}, \mathbf{p_2}, ...\mathbf{p_{c-1}}]$ and $\theta' = [\theta'_1, \theta'_2, ...\theta'_c]$, with the $p_{j's}$ being the mixing proportions and the $g_{j's}$, where $j = 1, 2, ..., c$ are the component densities, with density g_j being parameterized by θ_j [6]. C is the number of clusters or the components forming the mixture. The fundamental problem of the model based approach is the estimation of the parameters associated with the distribution we assumed for the population. Once that is settled, we can assign observations to particular clusters by maximizing a probability function. This is a posterior probability and it is of the form :

$$Pr(cluster\ j|x_i) = \frac{\hat{p}_j g_j(x_i, \hat{\theta}_j)}{f(x_j; \hat{p}, \hat{\theta})}\ j = 1, 2, ..., c$$

As stated by Everitt(1988), " in the case of finite mixture densities, the likelihood function is too complicated to employ the usual methods for its maximization, for example an iterative Newton–Raphson method which approximates the gradient vector of the log-likelihood function

$$l = \sum_{i=1}^{n} lnf(x_i; p, \theta)$$

by a linear Taylor series expansion" [5]. As a result, the iterative expectation maximization (EM) method or algorithm which was described in the work of Dempster et al.(1977) [15] has been widely used. There have been other Bayesian estimation methods that uses the Gibbs sampler or other Monte Carlo Markov chain (MCMC) methods which are becoming increasingly popular. Another algorithm for finite mixture is the classification maximum likelihood procedure, which was originally proposed by Scott and Symons (1971) [25].

2.2 Review of Previous Methodology for Clustering Mixed Data

There have been several scholarly works on the different algorithms for clustering mixed data. Most of the algorithms in earlier works focused on data in which some variables are continuous and some being categorical. One possibility would be doing a separate clustering on each variable type. This method is very popular but has one major drawback which is; conclusions from these separate analyses may not agree (Kaufman and Rousseeuw,1990) [3]. There is a second method of doing this which involves re-scaling all the variables to get them on a common scale. We then replace variable values by their ranks among the objects and then use a measure for continuous data [6]. Also, we can change all the variables in the data into binary form, or all into continuous variables. The problem with this is that treating variables as "what they are not" leads to loss of information. Kaufman and Rousseeuw(1990) stated this problem as follows: "Categorizing continuous variables via thresholds raises the question of what threshold value is appropriate and furthermore sacrifices a great deal of information" [3]. A third approach to clustering mixed data is based on the fact that clustering seeks to find similarities or dissimilarities based on distances and so we construct a dissimilarity measure for each type of variable and combine these into a single coefficient. One popular coefficient that has widely been used was proposed by Gower (1971) [4]. The coefficient for measuring similarity proposed by Gower made way for the clustering of mixed data with continuous, categorical and nominal variables. A generalization of the Gower coefficient was given by Kaufman and Rousseeuw to describe similarity as

$$1 - d(i, j)$$

where

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

and $\delta_{ij}^{(f)} = 1$ if the measurements x_{if} and x_{jf} for the fth variable are non-missing and 0 otherwise [3]. If f is binary or nominal then,

$$d_{ij}^{f} = \begin{cases} 1 & if \quad x_{if} \neq x_{jf} \\ 0 & if \quad x_{if} = x_{jf} \end{cases}$$

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

If the variable is interval scaled, then

$$d_{ij}^{f} = \frac{|x_{if} - x_{jf}|}{max_h(x_{hf}) - min_h(x_{hf})}.$$

Naturally, the use of the Gower coefficient comes with the likelihood of one variable gaining dominance. To remedy this problem, Chae, Kim and Yang (2006) [9] assigned weights to such variable types. In their work, they noted that assigning appropriate weights based on the characteristics of the data under consideration could overturn the dominance of one variable type. They defined a dissimilarity measure

$$d_{ij}^{*} = \tau_{ij} \sum_{l=1}^{c} \frac{1}{c} (\frac{|x_{il} - x_{jl}|}{R_l}) + (1 - \tau_{ij}) \sqrt{1 - A_{ij}},$$

with

$$A_{ij} = \frac{\sum_{l=c+1}^{r} s_{ijl}}{\sum_{l=c+1}^{r} w_{ijl}},$$

where $\tau_{ij}, 0 \leq \tau_{ij} \leq 1$, is a balancing weight such that

$$\tau_{ij} = \begin{cases} 1.0 - \frac{|\rho_{ij}^{c}|}{|\rho_{ij}^{c}| + |\rho_{ij}^{d}|} & \text{if } 1.0 < \frac{|\rho_{ij}^{c}|}{|\rho_{ij}^{d}|}, \\ 1.0 - \frac{|\rho_{ij}^{d}|}{|\rho_{ij}^{c}| + |\rho_{ij}^{d}|} & \text{if } 1.0 > \frac{|\rho_{ij}^{c}|}{|\rho_{ij}^{d}|}, \\ 0.5 & \text{if } |\rho_{ij}^{c}| = |\rho_{ij}^{d}|, \end{cases}$$

and $-1.0 \leq \rho_{ij}^{c}$ is the similarity measure for the quantitative variables, ρ_{ij}^{d} represents a similarity measure for the binary variables, $i = 2, 3, ..., n$ and $j = 1, 2, ..., n-1$, $i > j$. R_l is the range of the lth variable, $w_{ijl} = 1.0$ for continuous variables, $s_{ijl} = 1.0$ if $x_i = x_j$ and 0 otherwise, for binary variable, and w_{ijl} could either be 0 or 1, depending on whether the comparison between the ith and jth objects is valid for the lth variable. They used the

Pearson correlation coefficient in place of ρ_{ij}^{c} and also used the product moment correlation for ρ_{ij}^{d} [9] which is the Pearson correlation coefficient applied to binary data. Hendrickson (2014) gave an extension of the Gower coefficient [2] to include functional and directional variables. Her method will be discussed later under the simulation studies.

Clustering mixed data using finite mixture densities provides an optimal mathematically based approach to cluster analysis. Everitt (1988) proposed a clustering model which has been used over the years for clustering mixed data with continuous and ordinal or nominal variables. Several assumptions are made when this model is in use. It is assumed that an observed vector $\mathbf{x}$, consisting of $p+q$ random variables has a probability density function

$$f(\mathbf{x}) = \sum_{i}^{k} p_i MVN_{(p+q)}(\mu_i, \Sigma),$$

where k is the number of clusters, $p_1, p_2, ...p_k$ are the mixing proportions subject to the constraint $\sum_i^k p_i = 1$ and $MVN(.,.)$ denotes a multivariate normal density. Also, the binary and ordinal variables come from an underlying continuous distribution in which the q categorical or ordinal variables are generated by carefully setting some threshold values as cut-offs. The continuous variables $x_{p+1}, x_{p+2}, x_{p+3}, ..., x_{p+q}$ are observed only through categorized variables $z_1, z_2, z_3, ..., z_q$. The $z_j's$ are constructed in the following way,

$$z_j = \begin{cases} 1 & \text{if } -\infty = \alpha_{ij1} < x_{p+j} < \alpha_{ij2}, \\ 2 & \text{if } \alpha_{ij2} < x_{p+j} < \alpha_{ij3}, \\ t_j & \text{if } \alpha_{ijt_j} < x_{p+j} < \alpha_{ijt_j+1} = \infty \end{cases}$$

Where α_{ijl}, for $i = 1, ..., k$; $j = 1, ..., q$; and $l = 2, ..., t_j$ are the threshold values used to construct the ordinal variables, $z_1, z_2, z_3, ..., z_q$ from the continuous variables $x_{p+1}, x_{p+2}, x_{p+3}, ..., x_{p+q}$. He then proposed the density function

$$h(x, z) = \sum_{i=1}^{k} p_i MVN_{(p)}(\mu_i^{(p)}, \Sigma) \int_{a_1}^{b_1} ... \int_{a_q}^{b_q} MVN_q(\mu_i^{(q|p)}, \Sigma_{(q|p)}) dx_1, ...dx_q$$

where $\mu_i^{(q|p)} = \Sigma_{pq}' \Sigma_p^{-1}(x - \mu^{(p)})$ and $\Sigma_{q|p} = \Sigma_q - \Sigma_{pq}' \Sigma_p^{-1} \Sigma_{pq}$. It should be noted that these respectively, are the mean and covariance matrix for the conditional density $x_{p+1}, ..., x_{p+q}$

given $x_1, ..., x_{p+q}$. Σ_{pq} is the matrix of covariances between $x_1, ..., x_p$ and $x_{p+1}, ..., x_{p+q}$; Σ_P is the covariance matrix of $x_1, ..., x_p$; Σ_q is the covariance matrix of $x_{p+1}, ..., x_{p+q}$. As in the case of all other model based cluster analyses, the problem with clustering mixed data based on Everit's method is the estimation of the parameters associated with the distribution we assumed for the population. There is the need to estimate the parameters for a given set of observations in order to determine the probabilities for assigning the observations to appropriate clusters. In order to estimate the parameters, we maximize the log-likelihood function

$$logL = \Sigma_{i=1}^{k} g(x_i, z_i).$$

As we have discussed earlier in this work, maximizing this log-likelihood function requires the use of the the methods we discussed earlier under finite mixture densities (EM method or MCMC methods using the Gibbs sampler method).

3 PREVIOUS COMPARATIVE WORK AND PROPOSED WORK

3.1 Previous Comparative Work

Much attention has not been given to clustering mixed data as has been given to clustering the single data type. The few studies that have been conducted on clustering mixed data were done on mixed continuous and categorical variables (eg. The work of Alexander H. Foss and Marianthi Markatou) [17]. Traditionally, there are two widely used approaches to clustering data of mixed categorical and numeric attributes. One method involves the process of transforming categorical and nominal data into numeric integer values, and then applying numeric distance measures in order to compute similarity between object pairs. While this approach appears simple and straightforward, it has the disadvantage of assigning numeric values to categorical variables. Another approach has been to convert numerical attributes into discrete form and then applying a categorical clustering algorithm to it. The problem with this method is the loss of information due to the discretization process. Li and Biswas proposed a Similarity-Based Agglomerative Clustering (SBAC) algorithm that performs better in clustering data with mixed numeric and nominal features [21]. They adopted a similarity measure, proposed by Goodall (1966) for biological taxonomy, that gives greater weight to uncommon feature value matches in the computations of similarities and makes no assumptions about the underlying distributions of the feature values [22]. They used this to define the similarity measure between pairs of objects but their method is computationally expensive. There is another algorithm based on the K-means approach but this approach removes the numeric data only limitation whilst preserving the efficiency of the method. Under this, clustering is done with numeric and categorical attributes in a way similar to the K-means method but in this algorithm, objects are clustered against what is known as K-prototypes instead of the traditional K-means method of clustering and hence the name, K-prototypes algorithm [14]. In this method, we dynamically update the

K prototypes in order to maximize the intra-cluster similarity of objects from both numeric and categorical attributes [14]. The similarity measure used for the numeric attributes is the square euclidean distance whereas the similarity measure used on the categorical attributes is the number of mismatches between objects and cluster prototypes.

Hitherto, all the methods of clustering mixed data we have discussed have been on mixed numeric and categorical data types. Mixed data having categorical, continuous, directional and functional attributes have not been studied using clustering algorithms until recently. Notable among recent works on mixed data with all the above-mentioned attributes is the work by Hendrickson (2014). She used a model-based clustering method for mixed data based on Everitt's (1988) work. In Everitt's (1998) work, the main problem was estimating the parameters for the density h(·). These probabilities are what is actually used to identify the clusters. She suggested the use of a simulated annealing method to estimate the parameters for Everitt's model. In doing so, she used a penalized log likelihood with the simulated annealing method as a remedy for the parameter estimates being drawn to extremes [2]. She also used an extension of Gower's work on dissimilarities' measure to include functional and directional data types to estimate the Gower dissimilarity coefficient. The major task in her work was to include directional and functional data. In order to include directional variable, she used a dissimilarity measure described by Ackermann (1997) [11]. To measure the dissimilarity between two curves or functions, she used the L2 distance. The functional data she used had periodic signals.

3.2 Proposed Work

In this study, we will simulate data similar to that of Hendrickson (2014) but having a functional variable with: (a) strictly decreasing signal functions and (b) a mixture of periodic and strictly decreasing functions, to determine how the Extended Gower coefficient [2] performs on mixed data with directional variables and different signal functions (Strictly

decreasing, and decreasing with periodic and strictly decreasing tendencies). We will explore how the different hierarchical algorithms perform on mixed data with directional and functional variables and suggest an algorithm that works best under various settings. We will compare the clustering results of different hierarchical algorithms (complete linkage, average linkage, single linkage, and ward's method) when the functional data have no weights and how they perform when we attach weights to the functional data. We will use the inverse-variance weight as described by Chen et al. (2014) [10]. We will also examine how the various methods perform under various data scenarios.

4 SIMULATION STUDY

4.1 Setup of Study

This study was designed to test the performance of the Extended Gower coefficient (Hendrickson, 2014) [2] on mixed data that have directional variables and functional variables with different signal functions (Strictly decreasing, and decreasing with periodic and strictly decreasing tendencies). Four hierarchical clustering algorithms (single linkage, complete linkage, average linkage, and Ward's method) are implemented to determine how each performs on the data under different circumstances. The data simulated varied in the following ways: the number of objects in each true cluster, the different probability vectors for categorical variables, the different means and standard deviations for the continuous variables, the different values of the concentration variable kappa(κ) for the directional variable and the mean of the directional variable μ, the group of signal functions being used to generate the functional variable, and the standard deviation of the error that was added to the signal functions. In simulating the categorical data, we sampled observations with replacement from five categories based on the multinomial probability distribution function of the form

$$\frac{N!}{x_1!x_2!x_3!x_4!x_x!}p_1^{x_1}p_2^{x_2}p_3^{x_3}p_4^{x_4}p_5^{x_5},$$

where $N = \sum_1^5 x_i$, and p_i is the probability for each category [8].

We carefully selected two probability vectors similar to those used by Hendrickson [2]. For the first probability vector (0.8,0.05,0.05,0.05,0.05), the first entry represented a dominant category and the remaining four had the same values. The second vector of probabilities (0.2,0.2,0.2,0.2,0.2) represented equally likely categories. In doing this, we used the **R** function **sample.int** which has the following arguments: **n** which specifies the number of categories to choose from with **n=5** for this simulation, **size** for number of items to choose which we varied in different simulation settings; between **size=1,10,20,25,30,33 and 40**,

replace which specifies whether sampling is done with or without replacement and this was set to **TRUE** in this study, and **prob** representing the vector of probabilities. The continuous variable was simulated from a normal distribution with mean μ and standard deviation σ. In this study, σ was fixed at 100 for all four clusters and the mean μ was varied as follows: $\mu = 5000$ and for cluster 1, for cluster 2, $\mu = 5000 + k\sigma$, for cluster 3, $\mu = 5000 + 2k\sigma$, and for cluster 4, $\mu = 5000 + 3k\sigma$. The value of k was also chosen to vary from small to moderate to large. The following were chosen as values of **k**: $k = 5$, $k = 20$ and $k = 50$, which implied greater separation between clusters for greater values of **k**. The directional variable θ, was simulated by employing the von Mises distribution. The von Mises distribution is a continuous probability distribution with two parameters μ and κ. μ is the mean direction of the distribution, and κ is the concentration parameter of the distribution [12]. This distribution has density function:

$$\frac{exp(\kappa cos(\theta - \mu))}{2\pi I_0(\kappa)}, 0 \leq \theta \leq 2\theta,$$

where $0 \leq \mu < 2\pi$, $\kappa \geq 0$ and $I_0(\kappa)$ is the modified Bessel function defined by

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

[27]. In **R**, the **rvonmises** function in the **circular** package was used to simulate this distribution. For the 4 different clusters in this study, the following values were used for μ and κ: the value for κ was fixed at 50 for all 4 clusters and the value for μ was varied as follows; cluster 1, μ was 0, so that the data were highly concentrated around 0, for cluster 2, μ was $0+k$, for cluster 3, μ was $0+2k$, for cluster 4, μ was $0+3k$. Also the value of k varied from small to moderate to large; for example, we used $k = 0.5$, $k = 1.0$, and $k = 2.5$. The **rvonmises** generates random vectors following the von Mises distribution. The data can be spherical or hyper-spherical. The main difference between this study and other studies on clustering mixed data is the addition of directional and functional data. As a result, most

of the work focuses on the nature of the functional data. In simulating the functional data, we used two different types of signal functions. These signal functions are similar to those that were used by Ferreira and Hitchcock [7]. The first set of signal functions are of strictly decreasing tendencies. They were carefully chosen to reasonably lie close to one another so that the resulting clustering solution will be as good and reliable as possible, even with the addition of some noise (error). These functions are defined as follows:

$$\mu_1(t) = 50 - (t^2/500) - 7ln(t), t \in (0, 100]$$

$$\mu_2(t) = 50 - (t^2/500) - 5ln(t), t \in (0, 100]$$

$$\mu_3(t) = 50 - (t^2/750) - 7ln(t), t \in (0, 100]$$

$$\mu_4(t) = 50 - (t^2/250) - 9ln(t), t \in (0, 100]$$

These functions are plotted in Figure 4.1

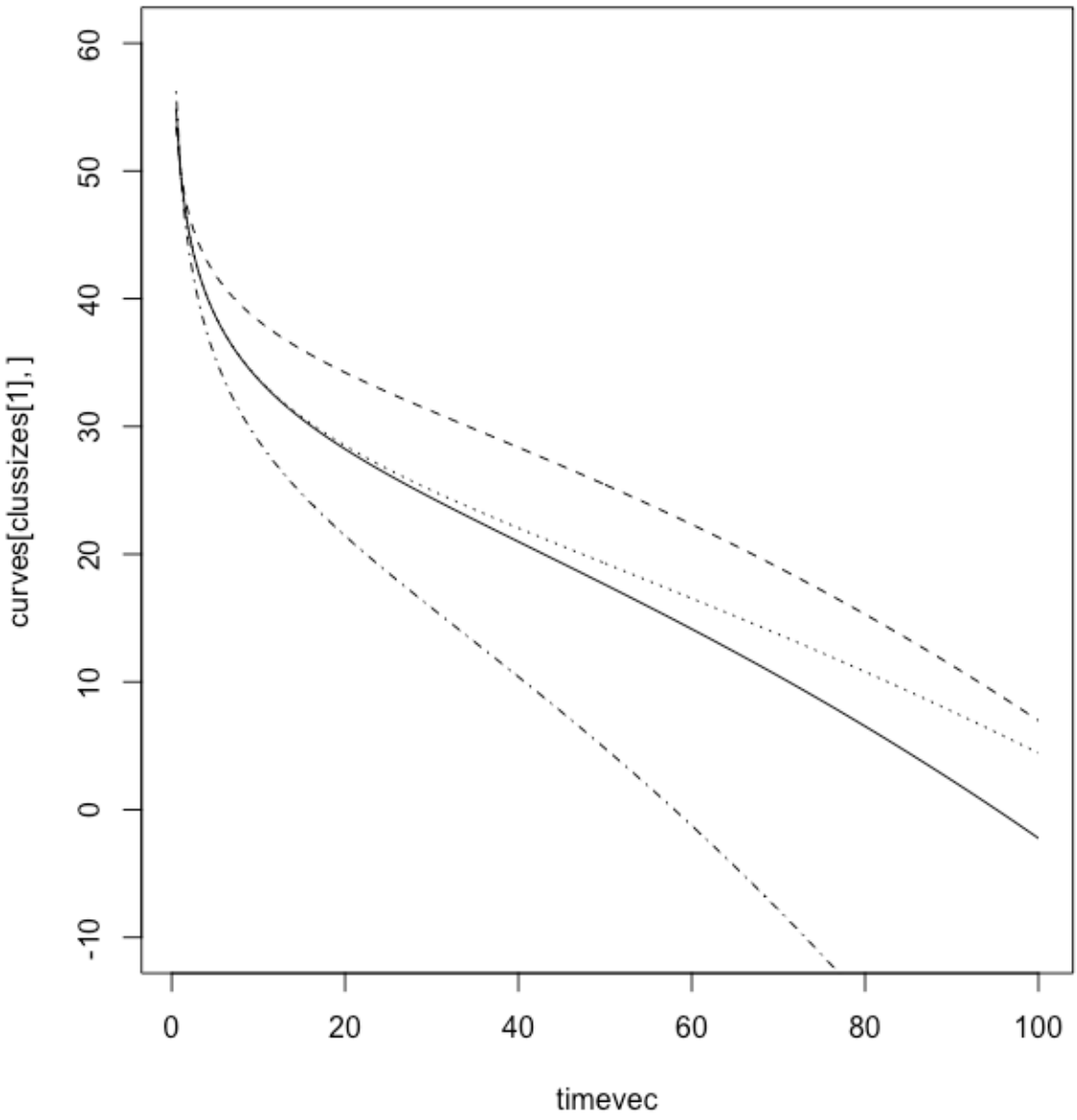

Figure 4.1: Signal Curves of The Strictly Decreasing Functions .

In all, 45 discretized curves were generated based on the four signal functions above to make up 45 different simulation settings. The data were simulated over 201 points from t=0.5 to t=200 in increments of 0.5. A random error was added to the signal functions based on the stationary Ornstein-Uhlenbeck process. A discretized approximation of the process such as the one used by Ferreira and Hitchcock [7] was used to produce an autoregressive covariance structure for the equally-spaced discretized data in the simulation. This process

is a Gaussian process with mean zero and the covariance between the errors measured at points t_i and t_j is $\sigma^2(2\beta)^{-1}exp(-\beta|t_i - t_j|)$ [7]. The drift variable, β, was kept at 0.5 and σ^2 was varied. The value of σ^2 was chosen for each group in a manner such as to test the robustness of the various clustering methods. Figure 4.2 shows the the addition of error (noise) using Ornstein-Uhlenbeck process with σ=1. Figure 4.3 shows the the addition of error (noise) when σ was chosen at σ=3 and Figure 4.4 is for σ=5

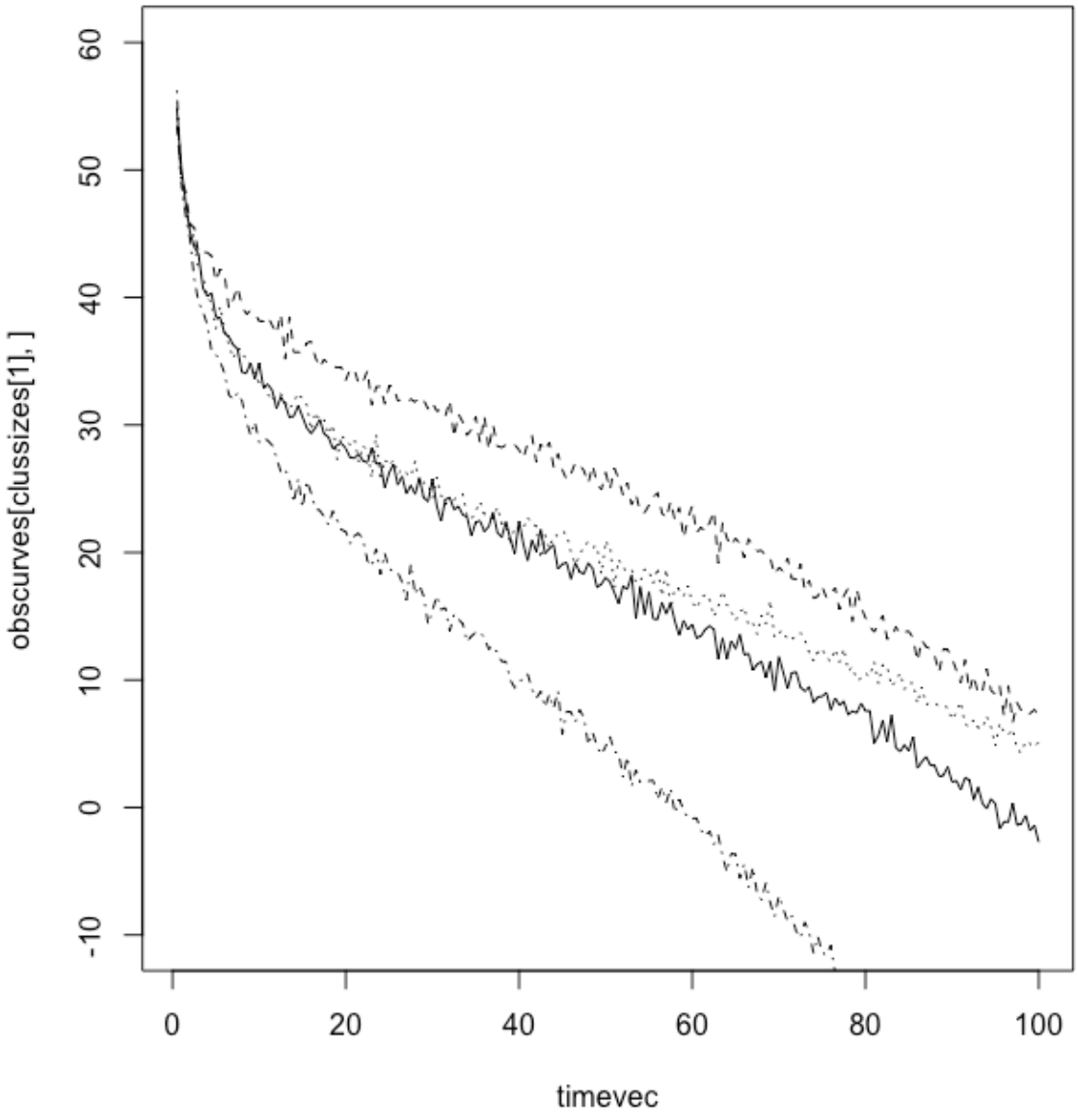

Figure 4.2: Strictly Decreasing Functions With Noise; σ=1 .

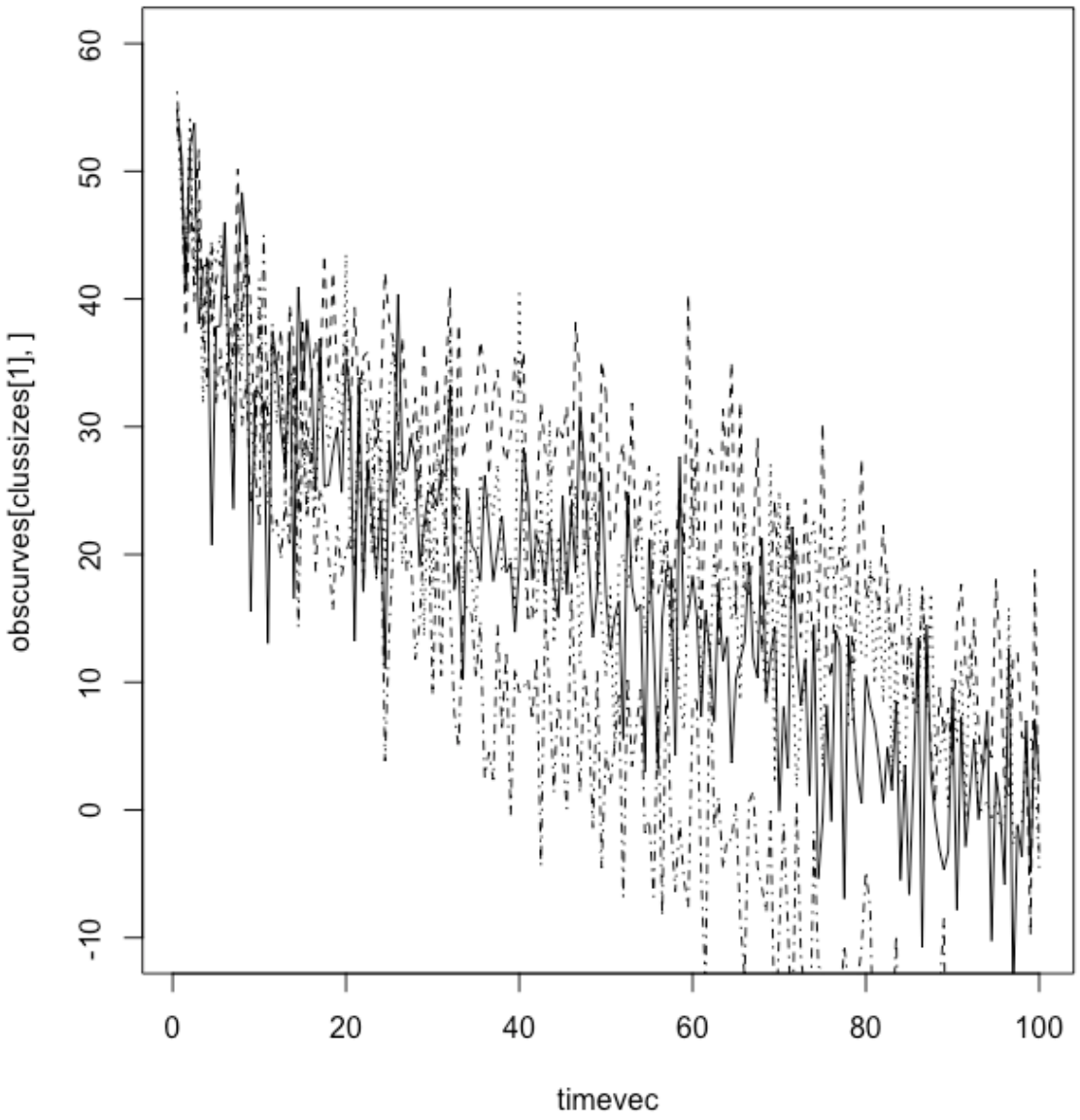

Figure 4.3: Strictly Decreasing Functions With Noise; σ=3 .

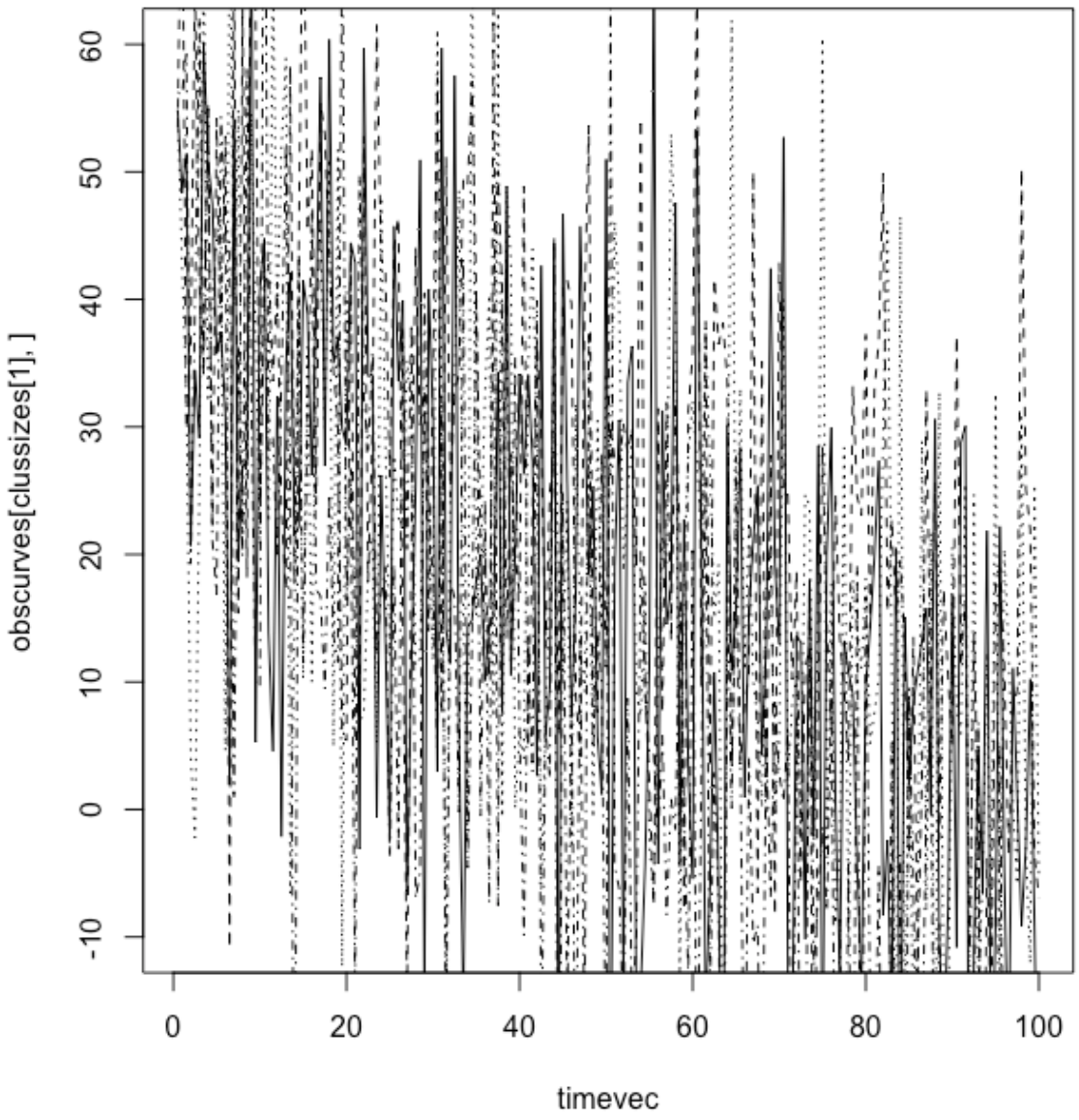

Figure 4.4: Strictly Decreasing Functions With Noise; σ=5 .

We now present the second group of signal curves. This group of curves had a decreasing trend, with a mixture of periodic and strictly decreasing functions. In all, 45 discretized curves were generated based on the four signal functions below to make up 45 different simulation settings. The simulation settings described for the strictly decreasing signal curves also applies to this group in the same way. The only difference is the form or the trend of

the signal curves. These signal curves are defined as follows:

$$\mu_1(t) = -t/2 + 2sin(t/5), t \in (0, 100]$$

$$\mu_2(t) = -t/2 + 2cos(t/3), t \in (0, 100]$$

$$\mu_3(t) = -t^2/250 - 4ln(t), t \in (0, 100]$$

$$\mu_3(t) = -t^2/250 - 2ln(t), t \in (0, 100]$$

These functions are plotted in Figure 4.5 below:

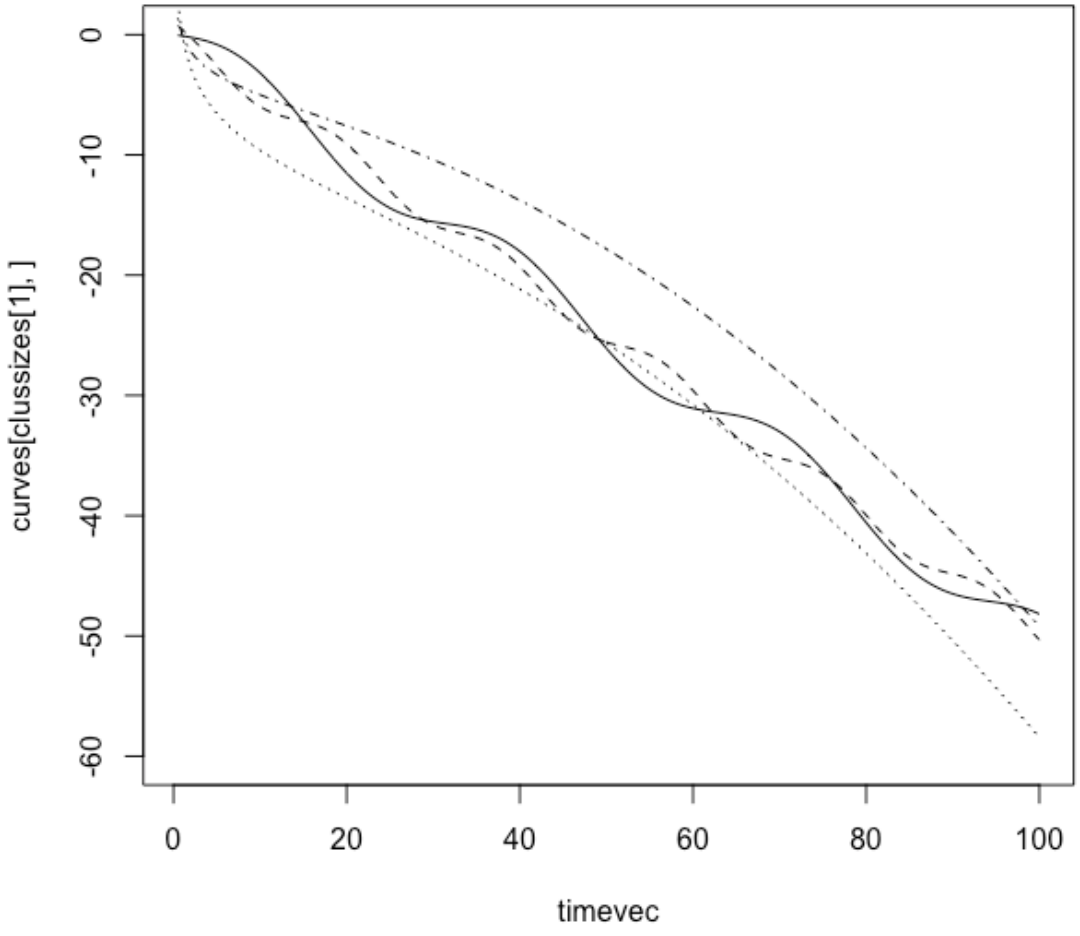

Figure 4.5: Signal Curves Of A Mixture Of Strictly Decreasing And Periodic Functions.

Figure 4.6 shows the the addition of error (noise) using Ornstein-Uhlenbeck process with σ=1. Figure 4.7 shows the addition of error (noise) for σ=3 and Figure 4.8 is for σ=5

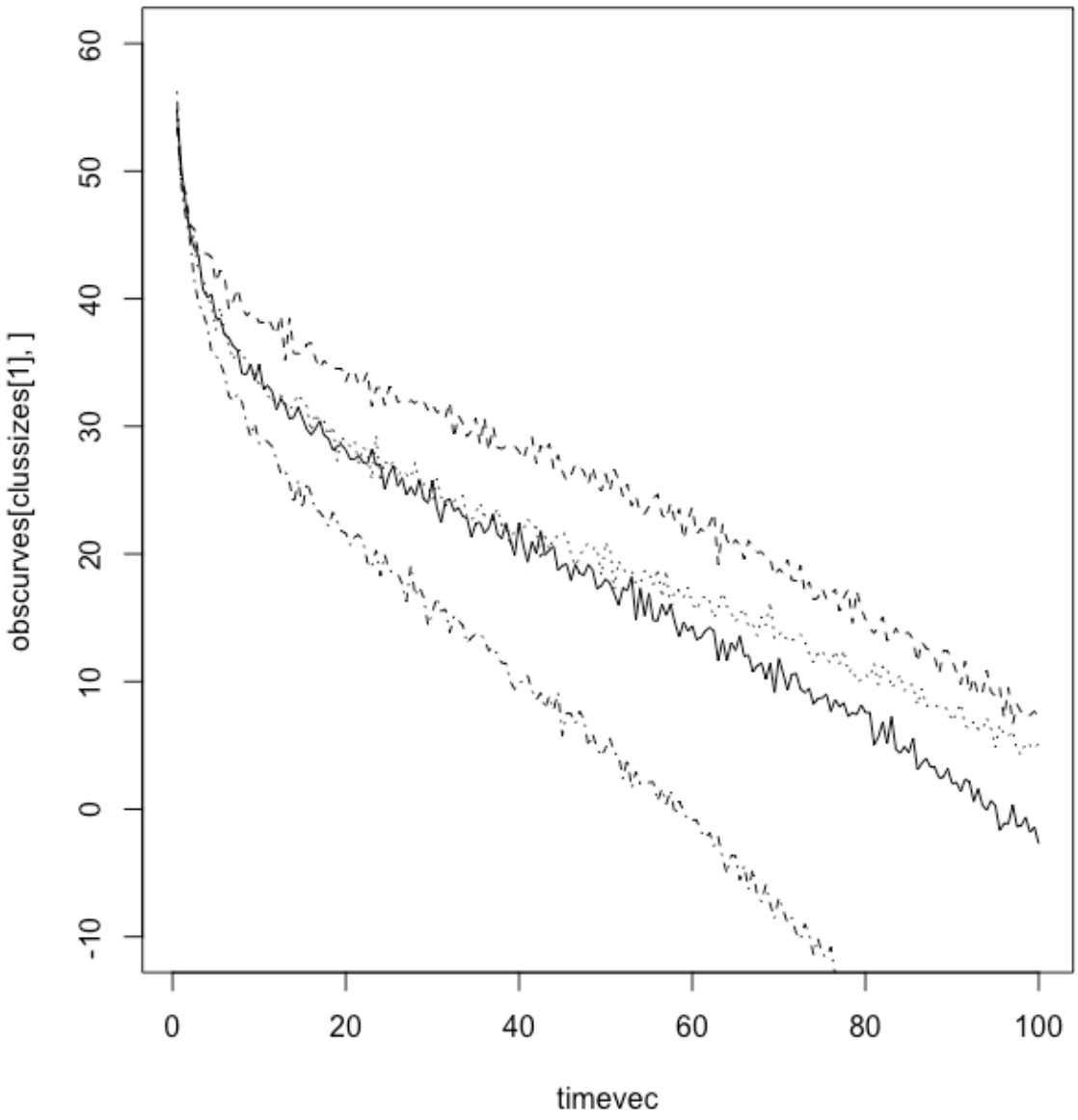

Figure 4.6: Signal Curves Of A Mixture Of Strictly Decreasing And Periodic Functions With Noise; σ=1 .

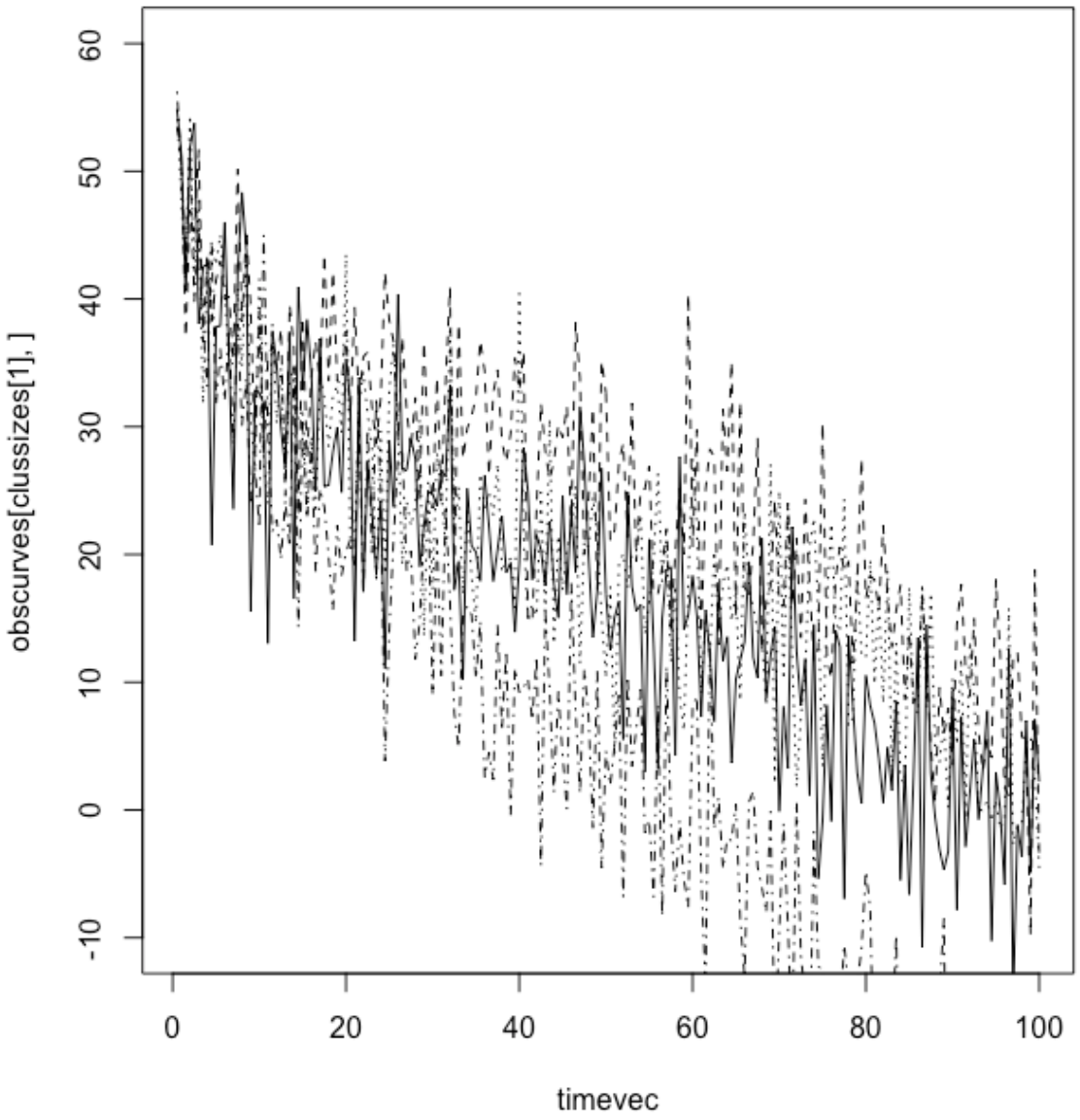

Figure 4.7: Signal Curves Of A Mixture Of Strictly Decreasing And Periodic Functions With Noise; σ=3 .

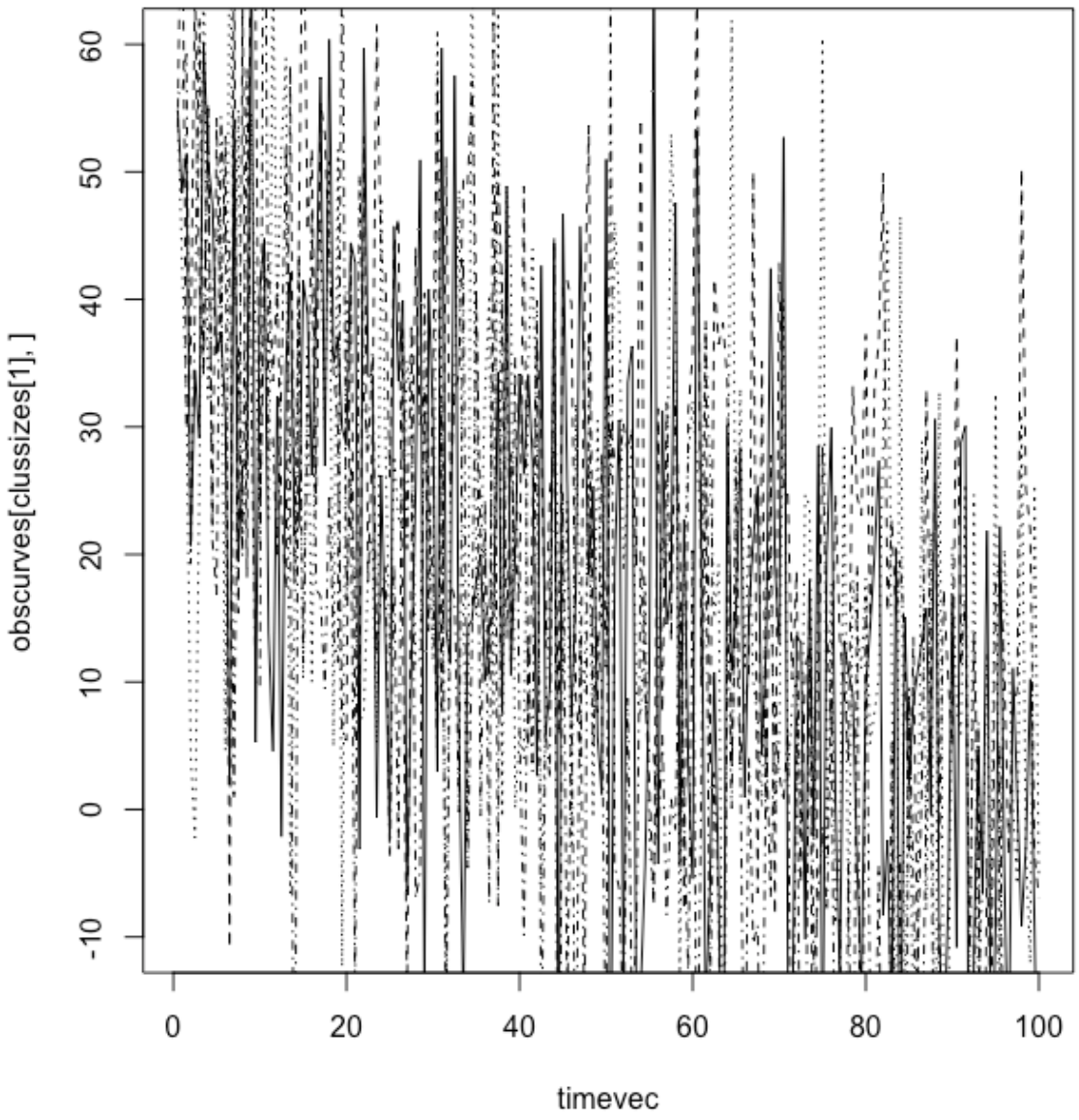

Figure 4.8: Signal Curves Of A Mixture Of Strictly Decreasing And Periodic Functions With Noise; σ=5 .

4.2 B-Splines

Naturally, our functional data should have a continuous functional structure but this is not the case in this situation and in many other instances. To demonstrate this, we can observe that the values of the functions we used in this study will converge to discrete data points once each of them is evaluated at $\mathbf{t}$, for all $\mathbf{t} \in (\mathbf{0}, \mathbf{100}]$. In order to convert each

discrete datum to a continuous functional observation, we used a method of smoothing called B-spline. By definition, B-splines are a combination of flexible bands that passes through the number of points that are called control points and creates smooth curves. B-splines gives a piecewise definition of what the best fit is, without changing the structure of the curves before clustering. We used the the **bs** function in the **splines** package of **R** to produce B-splines.

An extension of the Gower coefficient will allow for distances based on other types of variables, specifically functional and directional variables. In order to cluster the data using all of the variables jointly, we used the Gower extension.

4.3 The Extended Gower Coefficient

Hendrickson (2014) gave an extension of the Gower [4] coefficent to cluster mixed data with functional and directional variables. The dissimilarity between two objects i and j, is defined as follows

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

where $\delta_{ij}^{(f)} = 1$ if the measurements x_{if} and x_{jf} for the fth variable are non-missing and 0 otherwise (Kaufman Rousseeuw, 1990). The directional variable was added to the model using the measure of dissimilarity proposed by Ackermann (1997) [11]. He defined a similarity measure for directional variables as

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

where θ_i is the angle measured on object i. We wrote a function in R to compute this dissimilarity measure. To calculate the dissimilarities for the continuous or interval-scaled variables, we used the L_1 distance which defines the distance between observations x_i and x_j for the f^{th} variable as

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

The L_1 distance for the continuous variables in our simulation was computed in **R** using the **dist** function and setting the method as manhattan since the L_1 distance is similar to the manhattan distance. The dissimilarity component of the functional variable was computed using the the L_2 distance which is defined as

$$d_{L_2}^{(f)}(i,j) = \sqrt{\int_T [x_{if} - x_{jf}]^2 dt}.$$

This was executed in **R** using the **metric.lp** and **fdata** functions of the **fda.usc** package in **R**. The **fdata** function was used to convert the fitted values from the B-splines for each functional variable into a functional data object. This was done so that we could apply the **metric.lp** function since this function accepts only functional data objects. The **metric.lp** function computes an approximate L_2 distance for the functional data based on the Simpson's rule [23]

4.4 Weighted L_2 Distance

One objective of this study was to assess the impact of weights on the cluster solution. The essence of applying weights is to offset the effects of the dominance of one or more variables. Empirical research has shown that the functional variable dominates the clustering, for example as seen in the work of Hendrickson(2014) [2]. For this reason, we applied the inverse-variance weight function to the functional data, just as the work done by Oppong (2018) [29]. The inverse variance weight is defined as

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

where $\hat{\sigma}^2(t)$ is an estimate of the sample variance of all $y_i(t) - y_j(t)$ values for which $\sum_i w_i = 1$. This weight function applies more weight to areas of the curves with more spread, and less weight to curves which are less spread apart (Chen et al. 2014) [10]. This weight function

was then applied to the L_2 to get the weighted L_2 distance which is defined as;

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

To implement this in **R**, we used the **metric.lp** function with one additional argument, **w** for the weights. To use the extension of the Gower coefficient with our simulated data, we calculated the dissimilarities as has been described for each variable type. The dissimilarity between the ith and jth objects is the sum of all of the dissimilarities calculated for the ith and jth objects, divided by the sum of the number of variables, if we have that both measurements x_{if} and x_{jf} are for the fth variable and these variables are non-missing.

Table 4.1: Simulation Study Settings: Settings 1:8

Setting	Categorical Variable Probs.	Continuous Variable 1 Mean	Continuous Variable 2 Mean	Directional Variable Mean	Functional Variable sigma
1	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
2	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1$
3	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
4	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
5	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
6	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 3$
7	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 3$
8	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 3$

Table 4.2: Simulation Study Settings: Settings 9:15

Setting	Categorical Variable Probs.	Continuous Variable 1 Mean	Continuous Variable 2 Mean	Directional Variable Mean	Functional Variable sigma
9	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 3$
10	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 3$
11	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 5$
12	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 5$
13	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 5$
14	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 5$
15	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 5$

As can be seen in Table 4.1 and Table 4.2, the simulation settings were varied in many ways. This was to allow us to do an assessment of the Extended Gower coefficient [2] under the various settings. The first variation we introduced was the cluster sizes. In all, we had four clusters for each simulation setting. The cluster sizes were chosen carefully to represent the following scenarios:

- equal cluster sizes: 25 objects in each of the four clusters
- 3 equal-sized clusters and 1 smaller cluster: 33 objects in the first three clusters and 1 object in the last cluster
- different cluster sizes with a common difference: 10, 20, 30, 40.

All other variations apply as has been discussed in the earlier part of this section. For each combination of cluster-size, categorical variable generation scheme, continuous variables' parameterization, and different signal functions, 1000 datasets were generated. In all, we had 15000 data sets for mixed data whose signal functions had a strictly decreasing tendency, and 15000 datasets for mixed data whose signal functions had a mixture of strictly decreasing and periodic tendencies. Based on the 1000 iterations for each setting, the mean Rand index and Monte Carlo standard error for the mean Rand index was computed to be used for performance comparison of the different clustering algorithms. The same thing was repeated again for all the simulation settings with an addition of weights to the functional data in order to assess how weights affects the clustering solution.

4.5 Dendogram

A dendogram is a tree structured graph used to visualize the results of a hierarchical clustering, by portraying the hierarchical relationship between the objects in the clustering results. To determine the similarity between two objects on a dendogram, we consider the

height of the link that joins them. Similar objects have a shorter link and thus belong to a common cluster.

4.6 The Rand Index

Rand (1971), defined an index for validating cluster solutions. It is a concordance index to compare two different clustering solutions say C_1 and C_2 based on pairs of objects in the data. The Rand index is defined as

$$R = \frac{N_{11} + N_{00}}{N_{00} + N_{01} + N_{10} + N_{11}}$$

where N_{11} is the number of pairs of objects correctly placed in the same cluster by both clustering, N_{00} is the number of pairs of objects that are correctly placed in different clusters, N_{10} is the number of pairs of objects that are in the same cluster in C_1, but in different clusters in C_2, N_{01} is the number of pairs of objects that are in different clusters in C_1, but in the same clusters in C_2. This results in a value of R with: $0 \leq R \leq 1$, where values close to 0 indicates that the two data clustering do not agree and a value close to 1 indicates that data are clustered in nearly the same way. Other methods of cluster performance assessment is the adjusted Rand index which is a version of the Rand index adjusted for agreement due to chance. Cohen's (1960) [24] κ statistic is another measure. For simplicity and consistency with similar studies, we will use the Rand index.

4.7 The Monte Carlo Standard Error(MCSE)

The MCSE measures the variability of the Rand index across simulation iterations [7]. It suffices to compare two or more mean Rand values just as they are given by the results of our clustering algorithm if they differ by more than, say, twice the associated MCSE. In this study, the MCSEs are very small so we will loosely compare the mean Rand values without considering the vagaries or unpredictability of the simulated data sets. We will however present their values.

4.8 Results

Tables 3 to 10 of appendix A give the average Rand indices of the results from our simulations. For the mixed data with strictly decreasing functional variables (see Table 7 and Table 8), the extended Gower coefficient generally performed well under all the various clustering algorithms. In almost all simulation settings under this type of data, the values of the average Rand indices were close to 1, with some of them actually getting to 1, except for cases when the single linkage method was used. The single linkage method performed the least in almost all of the simulation settings. The single linkage method produced the worst Rand index values of 0.3376 for simulation setting 11C, 0.3535 for simulation setting 9B and 5B, and 0.3810 for simulation setting 2B. Among the remaining clustering algorithms, it was hard to select which method performed the best since all of them produced Rand indices which were close to 1 and even 1 in some cases. For instance in simulation setting 2C, the Ward's method produced a Rand index of 1, followed by the complete linkage method with a rand index of 0.9489 and then the average linkage method with a mean Rand value of 0.9461. In simulation setting 3b, the average linkage method also gave an average Rand index of 1, followed by the complete linkage method with a mean Rand value of 0.9741 and then the Ward's method with a mean Rand value of 0.9612. Hitherto, we have been discussing the general performance of the methods. In order to make a comparison as to the most suitable method for a particular situation based on the results of our simulations, we will consider the simulation settings under which our favorite three clustering methods (Complete, Average, Ward) performed "moderately" well. That is, situations under which they produced "small" mean Rand values. We will select the most suitable clustering algorithm based on their performance under those "special" scenarios. Simulation setting 12b produced the "least" mean Rand value among all the mean Rand values of the three methods. The complete linkage had the "highest" mean Rand value of 0.7014 under that setting, followed by the

Ward's method with a mean Rand value of 0.5804 and then the average linkage method with a mean Rand value of 0.5725. Under this setting, there was a large separation between the clusters for the directional and functional variables, and a small separation between the clusters for the continuous and categorical variables. The value of σ was 5 and the cluster sizes were: 33, 33, 33, and 1. Figure 4.9 shows the dendogram for the complete linkage method of clustering mixed data with strictly decreasing functional data component. The dendogram shows that the clusters are merged at shorter distances, and the clusters are also well separated.

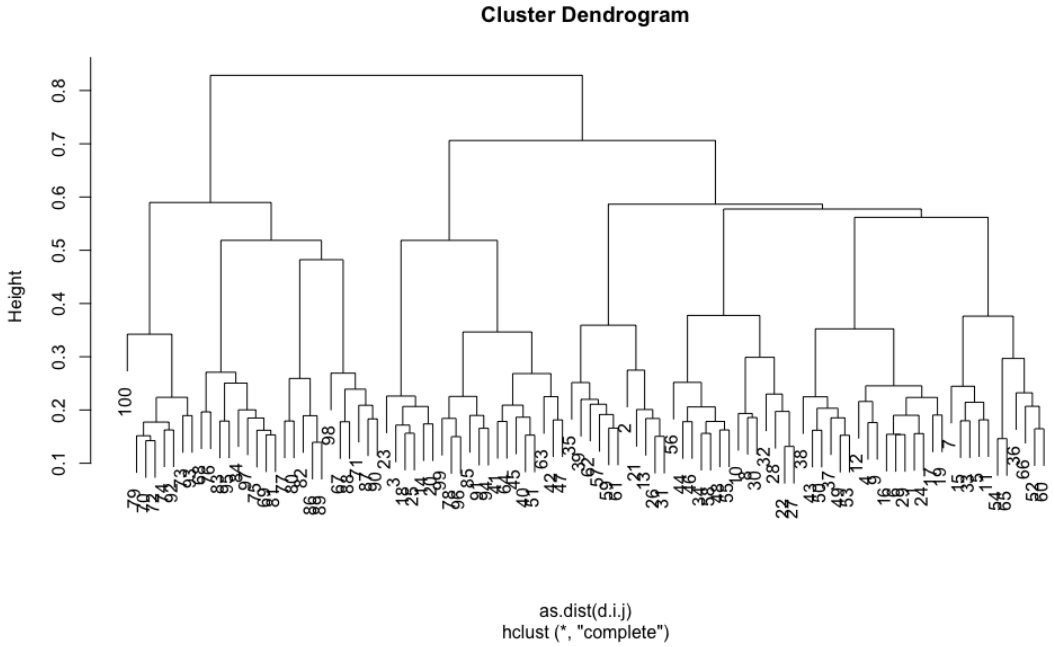

Figure 4.9: Dendrogram-Complete linkage Method for Mixed Data With Strictly Decreasing Functional Data

Also in simulation settings 12a, 12c, and 13a, where there were large separations between the clusters for the directional and functional variables, and small separations between the clusters for the continuous and categorical variables with the cluster sizes cutting across all the different types cluster sizes used in this study, the complete linkage method performed better than all the other methods, followed by the Ward's method and then the average

linkage method. Based on these results, the complete linkage method is the best method to use to cluster mixed data with strictly decreasing functional variables when there is a large separation between the clusters for the directional and functional variables, and small separation between the clusters for the continuous and categorical variables.

When the inverse-variance weight function was applied to this type of mixed data, we generally observed an improvement in the mean Rand values (see table 9 and 10 of appendix A). We averaged all the mean Rand values observed across simulations before applying weights to the functional data and we had an average value of 0.8278. This average value increased to 0.8313 after the application of weights. In most cases, the mean Rand values remained the same or increased but there were no recognizable patterns associated to those occurrences. For this reason, we generalized that the weights improved the cluster solution. Also, there were notable circumstances when the mean Rand values reduced after the application of weights to the functional data. For instance in simulation 6c, the mean Rand value for the complete linkage method reduced from 0.8994 when there was no weight to 0.7630 when the inverse variance weight was applied. Also, the mean Rand value for the average linkage method reduced from 0.8709 when there was no weight to 0.7630 when the weight was applied, and the mean Rand value for the single linkage method reduced from 0.8305 to 0.6487 after applying the weights. Other simulation settings in which we had the mean Rand values reducing after weight application are simulation setting 7A, 7C, 8A, 9A, 9C, 10C, 13A, and 13C. All these occurred in situations where the clusters of all the different data types were largely well separated.

Just like the mixed data with strictly decreasing functional component, the extended Gower coefficient did well on the mixed data whose functional component had a mixture of strictly decreasing and periodic signal curves (See table 3 and 4 of appendix A). Under this mixed data type, the single linkage method performed poorly as compared to the remaining

three methods. For instance, if we look at the results from simulation 1b, 4b, and 6b, while the complete linkage, average linkage and Ward's method produced a mean Rand value of 0.9869, 0.9901, and 0.9489 respectively, the single linkage method produced mean Rand values of 0.3606, 0.3606, and 0.5897 in those respective settings. The single linkage method did not perform well across simulations but its performance was worse when the cluster sizes were 33, 33,33, 1. Under this situation, there was a large separation between the clusters for the directional and functional variables, and a small separation between the clusters for the continuous and categorical variables. For the remaining three methods, it was hard to choose the one which performed best when the standard deviation of the functional data (noise component) σ was relatively small. Thus all the other methods performed almost the same when there were small separations between the clusters of all the different data types. The mean Rand values obtained in most simulations under those circumstances however placed the complete linkage and Ward's method above the average linkage method. As a suggestion, choosing between the complete linkage and ward method should be based on the structure of the data you are dealing with and how easier or comfortable the usage of any of those two methods is to the user, as well as the run-time of the algorithm. However when σ was increased to 5, the complete linkage method performed better than the Ward's method except for special cases when the cluster sizes were 33, 33, 33, 1 and there was a large separation between the clusters for the directional and functional variables, and small separation between the clusters for the continuous and categorical variables.

When the inverse-variance weight function was applied to this type of mixed data (see Table 5 and 6 of appendix A), we generally observed a decrease in the mean Rand values. We averaged all the mean rand values observed across simulations before applying weights to the functional data and we had an average value of 0.8046 which reduced to 0.8005 after the application of weights. In most cases, the mean rand values remained the same or decreased

after the application of weights but there was no recognizable pattern associated to those occurrences. For this reason, we generalize that the weights did not improve the cluster solution.

In general, the complete linkage method performs the best since it produced the highest mean Rand value in most situations except for situations where the data contained one very small group and a few large groups, with a large σ value (more noise), in which case the Ward's method performed better. Although the average linkage method performed quite well, it mostly ranked after the complete linkage and Ward's method. The single linkage method performed worst overall, except in some few cases. For these reasons, we recommend using the complete linkage for cases when the data structure or the natural clusters in the data set are unknown. Also, when the analyst suspects or have strong evidence of the presence of much noise in the mixed data, we recommend using the Ward's method if there is also an evidence of the existence of one or two extremely large or extremely small groups in the dataset, and there exist a large separation between the clusters for the directional and functional variables and a small separation between the clusters for the continuous and categorical variables. The choice of using the Ward's method or Complete average method should be based on computational flexibility and run-time when there is reason to believe that the mixed data has small noise or small separation between the clusters of all the different data types. The single linkage method should never be used as it performs poorly in general. All these apply in the two types of mixed data we have studied. Also, the inverse variance weight improved the clustering solution for the mixed data with strictly decreasing functional components but did not improve the clustering solution of the mixed data with a mixture of strictly decreasing and periodic signal functional component.

5 DISCUSSION / FUTURE RESEARCH

We looked at the extension of the Gower's coefficient to include directional and functional data by Hendrickson(2014). However, we used two different functional data types hence, our study was on two different types of mixed data; one with a functional component whose signal curves had strictly decreasing tendencies and the other one with functional component whose signal curves had a mixture of strictly decreasing and periodic tendencies as in the work of Hitchcock and Feirrera(2009). To approach these, we simulated 45 datasets for each type of mixed data we studied. For each simulation, we had 1000 iterations. We introduced some noise into the datasets using the stationary Ornstein-Uhlenbeck process whose effects depended on the choice of σ we used. In this study, we used 3 different σ values to reflect different data situations. We also used different cluster sample sizes to reflect different data structures in order to determine how our methods performed under all the different settings. We sort to determine how well the extension of the Gower coefficient performed on the two types of mixed data. We did a performance comparison of four different agglomeraitve hierarchical clustering algorithms(complete linkage,average linkage, single linkage and the Ward's method). We also applied the inverse variance weight function to the functional data in both cases to determine its impact on our solutions. To do all the assessments and comparison, we computed the mean Rand Indices of the various simulations. In general, the extension of the Gower coefficient performed well under almost all the simulation settings. The Complete linkage method and Ward's method were found to be the dominant methods but then the complete linkage was more effective except for situations where the data contained one very small group and a few large groups with a large separation between the clusters for the directional and functional variables, and small separation between the clusters for the continuous and categorical variables, in which case the Ward's method performed well. The performance of the Ward's method is consistent with Hitchcock and Fereiera's(2009) [7] finding. Though

their work was on functional data, the dominance of the functional data in a mixed data as established in the results of Hendrickson(2014) [2] makes this a plausible conclusion. The single linkage method was the worst method in almost all the simulation settings. Applying the inverse variance weights improved the clustering solution for the mixed data with strictly decreasing functional curves but did not improve the clustering solution of the other type of mixed data. We recommend using the complete linkage for cases when the data structure or the natural clusters in the dataset are unknown. Also, when the analyst suspects or have strong evidence of the presence of much noise and there is also evidence of the existence of one or two extremely large or extremely small groups in the mixed data, we recommend using the Ward's method. Such decisions should also consider computational flexibility and the method's suitability with the data structure as well as the run-time of the algorithm. The single linkage method should not be used under any circumstance and if there is the need to apply weights, other potential weight functions (not the inverse-variance) should be explored when the data has a mixture of strictly decreasing and periodic functional components. As an option for future research and development of the methods explored in this work, we suggest that different weight functions be applied to the functional data. Also, a method for introducing dependencies in the simulation of the different data types should be explored.

REFERENCES

[1] Gordon,A.D, *Classificaiton-Monographs on applied probability and statistics*. ISBN 0-412-22850

[2] Hendrickson, J. L.(2014). Methods for Clustering Mixed Data. (Doctoral dissertation). Retrieved from `https://scholarcommons.sc.edu/etd/2590`

[3] Leonard Kaufman and Peter J. Rousseeuw. *Finding groups in data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley and Sons Inc., New York, 1990. An introduction to cluster analysis, A Wiley-Interscience Publication.

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

[5] B. S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statist. Probab. Lett.*,6(5):305–309, 1988.,

[6] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Analysis, volume 848 of Wiley series in probability and statistics*. John Wiley and Sons, 2011.

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

[8] E.L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer Texts in Statistics. Springer, 1998.

[9] Kim Jong-Min Chae, Seong San and Wan Youn Yang. Cluster analysis with balancing weight on mixed-type data. *The Korean Communications in Statistics*, 13(3):719–732, 2006.

[10] Chen Huaihou, Reiss Philip T. and Tarpey Thaddeus. Optimally weighted L2 distance for functional data. *Biometrics* volume 70 number = 3,issn = 1541-0420 Retrieved from `http://dx.doi.org/10.1111/biom.12161`

[11] H. Ackermann. A note on circular nonparametrical classification. Biometrical Journal, 5:577–587, 1997.

[12] Claudio Agostinelli and Ulric Lund. circular: Circular Statistics, 2011. R package version 0.4-3.

[13] Johnson, R.A. and Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River.

[14] Huang, Z. (1998). Extensions to the k-means algorithms for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, Vol. 2, 283-304.

[15] Dempster, A., Laird, N., Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1-38. Retrieved from `http://www.jstor.org/stable/2984875`

[16] P. MacNaughton-Smith, W. T. Williams, M. B. Dale L. G. Mockett .Dissimilarity Analysis: a new Technique of Hierarchical Sub-division Nature volume 202, pages 1034–1035 (06 June 1964)

[17] Alexander H. Foss ,Marianthi Markatou. Clustering Mixed-Type Data in R and Hadoop Journal of Statistical Software February 2018, Volume 83, Issue 13. doi: 10.18637/jss.v083.i13

[18] Ward, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236-244.

[19] Sorensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter, 5:1-34.

[20] MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281–297, University of California Press, Berkeley, Calif., 1967. Retrieved from `https://projecteuclid.org/euclid.bsmsp/1200512992`

[21] C. Li and G. Biswas.Unsupervised learning with mixed numeric and nominal data.IEEE Transactions on Knowledge and Data Engineering (Volume: 14 , Issue: 4 , Jul/Aug 2002)

[22] D. W. Goodall. A new similarity index based on probability. Biometrics, 22(4):882–907, 1966.

[23] Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The r package fda.usc. *Journal of Statistical Software*, 51(4), 2012.

[24] Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 37-46.

[25] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971.

[26] Rencher A.C Methods of Multivariate Analysis Volume 492 of Wiley Series in Probability and Statistics

[27] S.R. Jammalamadaka and A. Sengupta. *Topics in Circular Statistics.* Series on multivariate analysis. World Scientific, 2001

[28] Sneath, P. H. A. (1957). The application of computers to taxonomy. Journal of General Microbiology, 17, 201– 226.

[29] Oppong, Augustine, "Clustering Mixed Data: An Extension of the Gower Coefficient with Weighted L2 Distance" (2018). Electronic Theses and Dissertations. Paper 3463. Retrieved from `https://dc.etsu.edu/etd/3463`

APPENDIX

RAND INDICES AND MONTE CARLO STANDARD ERRORS

Table .1: Simulation Setting 1A-8C: Average Rand Index (MCSE) For Mixed Data With a Mixture of Strictly Decreasing and Periodic Functions-unweighted.

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
1A	0.9901 (0.0010)	0.9901 (0.0010)	0.8644 (0.0009)	0.9901 (0.0010)
1B	0.9869 (0.0010)	0.9869 (0.0010)	0.3606 (0.0004)	0.9489 (0.0009)
1C	0.9941 (0.0010)	0.9941 (0.0010)	0.8701 (0.0009)	1.0000 (0.0010)
2A	0.8600 (0.0009)	0.9622 (0.0010)	0.6216 (0.0006)	0.9519 (0.0010)
2B	0.8554 (0.0009)	0.8709 (0.0009)	0.7543 (0.0008)	0.8798 (0.0009)
2C	0.9501 (0.0010)	0.9887 (0.0010)	0.9887 (0.0010)	0.9784(0.0010)
3A	0.9545 (0.0010)	0.9545 (0.0010)	0.8604 (0.0009)	0.9467 (0.0009)
3B	0.9618 (0.0010)	0.9618 (0.0010)	0.6293 (0.0006)	0.9448 (0.0009)
3C	0.9887 (0.0010)	0.9887 (0.0010)	0.8376 (0.0008)	0.9887 0.0010
4A	0.9901 (0.0010)	0.9622 (0.0010)	0.8402 (0.0008)	0.9622 (0.0010)
4B	0.9741 (0.0010)	0.9117 (0.0009)	0.3606 (0.0004)	0.9105 (0.0009)
4C	0.9790 (0.0010)	0.9570 (0.0010)	0.8378 (0.0008)	0.9790 (0.0010)
5A	1.0000 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)
5B	1.0000 (0.0010)	0.9869 (0.0010)	0.7549 (0.0008)	0.9545 (0.0010)
5C	1.0000 (0.0010)	1.0000 (0.0010)	0.8770 (0.0009)	1.0000 (0.0010)
6A	0.7489 (0.0007)	0.7461 (0.0007)	0.6269 (0.0006)	0.8422 (0.0008)
6B	0.7168 (0.0007)	0.7364 (0.0007)	0.5897 (0.0006)	0.5962 (0.0006)
6C	0.7980 (0.0008)	0.7735 (0.0008)	0.5958 (0.0006)	0.7129 (0.0007)
7A	0.8242 (0.0008)	0.8745 (0.0009)	0.5327 (0.0005)	0.8669 (0.0009)
7B	0.7174 (0.0007)	0.7489 (0.0007)	0.6604 (0.0007)	0.8321 (0.0008)
7C	0.7943 (0.0008)	0.9038 (0.0009)	0.8232 (0.0008)	0.8851 (0.0009)
8A	0.9901 (0.0010)	0.9014 (0.0009)	0.6497 (0.0006)	0.9014 (0.0009)
8B	0.9279 (0.0009)	0.8547 (0.0009)	0.4487 (0.0004)	0.9388 (0.0009)
8C	0.9628 (0.0010)	0.8887 (0.0009)	0.3772 (0.0004)	0.9715 (0.0010)

Table .2: Simulation Setting 9A-15C: Average Rand Index (MCSE) For Mixed Data With a Mixture of Strictly Decreasing and Periodic Functions-unweighted.

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
9A	0.9206 (0.0009)	0.8588 (0.0009)	0.3103 (0.0003)	0.8741 (0.0009)
9B	0.9143 (0.0009)	0.8796 (0.0009)	0.8042 (0.0008)	0.8709 (0.0009)
9C	0.9325 (0.0009)	0.9053 (0.0009)	0.4871 (0.0005)	0.9481 (0.0009)
10A	0.8168 (0.0008)	0.7865 (0.0008)	0.6180 (0.0006)	0.8659 (0.0009)
10B	0.7832 (0.0008)	0.7802 (0.0008)	0.5582 (0.0006)	0.7776 (0.0008)
10C	0.7671 (0.0008)	0.7731 (0.0008)	0.5291 (0.0005)	0.8620 (0.0009)
11A	0.9261 (0.0009)	0.9261 (0.0009)	0.5828 (0.0006)	0.9244 (0.0009)
11B	0.8644 (0.0009)	0.8960 (0.0009)	0.4018 (0.0004)	0.8644 (0.0009)
11C	0.7554 (0.0008)	0.8834 (0.0009)	0.3416 (0.0003)	0.8729 (0.0009)
12A	0.7507 (0.0008)	0.6010 (0.0006)	0.6055 (0.0006)	0.6463 (0.0006)
12B	0.7220 (0.0007)	0.5725 (0.0006)	0.5723 (0.0006)	0.5804 (0.0006)
12C	0.7063 (0.0007)	0.6354 (0.0006)	0.5659 (0.0006)	0.6002 (0.0006)
13A	0.7451 (0.0007)	0.6818 (0.0007)	0.6087 (0.0006)	0.6818 (0.0007)
13B	0.7208 (0.0007)	0.6996 (0.0007)	0.5715 (0.0006)	0.7042 (0.0007)
13C	0.7218 (0.0007)	0.7036 (0.0007)	0.6067 (0.0006)	0.7097 (0.0007)
14A	0.8657 (0.0009)	0.8830 (0.0009)	0.7206 (0.0007)	0.8477 (0.0008)
14B	0.9182 (0.0009)	0.8776 (0.0009)	0.6733 (0.0007)	0.9281 (0.0009)
14C	0.9107 (0.0009)	0.8996 (0.0009)	0.6810 (0.0007)	0.8513 (0.0009)
15A	0.8673 (0.0009)	0.7780 (0.0008)	0.6667 (0.0007)	0.7885 (0.0008)
15B	0.7299 (0.0007)	0.7404 (0.0007)	0.5564 (0.0006)	0.7525 (0.0008)
15C	0.7053 (0.0007)	0.7921 (0.0008)	0.5143 (0.0005)	0.6758 (0.0007)

Table .3: Simulation Setting 1A-10C : Average Rand Index (MCSE) For Mixed Data With a Mixture of Strictly Decreasing and Periodic Functions(weighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
1A	0.9901 (0.0010)	0.9901 (0.0010)	0.8644 (0.0009)	0.9901 (0.0010)
1B	0.9869 (0.0010)	0.9869 (0.0010)	0.3606 (0.0004)	0.9489 (0.0009)
1C	0.9941 (0.0010)	0.9941 (0.0010)	0.9887 (0.0010)	1.0000 (0.0010)
2A	0.8642 (0.0009)	0.9622 (0.0010)	0.6366 (0.0006)	0.9519 (0.0010)
2B	0.8420 (0.0008)	0.8160 (0.0008)	0.6186 (0.0006)	0.7976 (0.0008)
2C	0.9887 (0.0010)	0.9887 (0.0010)	0.9887 (0.0010)	0.9784 (0.0010)
3A	0.9545 (0.0010)	0.9545 (0.0010)	0.8446 (0.0008)	1.0000 (0.0010)
3B	0.9618 (0.0010)	0.9618 (0.0010)	0.5444 (0.0005)	0.9448 (0.0009)
3C	0.9887 (0.0010)	0.9887 (0.0010)	0.8376 (0.0008)	0.9887 (0.0010)
4A	0.9901 (0.0010)	0.9901 (0.0010)	0.8402 (0.0008)	0.9622 (0.0010)
4B	0.9257 (0.0009)	0.8994 (0.0009)	0.3731 (0.0004)	0.9248 (0.0009)
4C	0.9790 (0.0010)	0.9790 (0.0010)	0.8578 (0.0009)	0.9790 (0.0010)
5A	1.0000 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)
5B	0.9869 (0.0010)	0.9869 (0.0010)	0.3675 (0.0004)	0.9545 (0.0010)
5C	1.0000 (0.0010)	1.0000 (0.0010)	1.0000 (0.0010)	1.0000 (0.0010)
6A	0.7489 (0.0007)	0.7461 (0.0007)	0.6269 (0.0006)	0.8422 (0.0008)
6B	0.7168 (0.0007)	0.7364 (0.0007)	0.5897 (0.0006)	0.5962 (0.0006)
6C	0.7735 (0.0008)	0.7143 (0.0007)	0.5958 (0.0006)	0.7129 (0.0007)
7A	0.8242 (0.0008)	0.8745 (0.0009)	0.5327 (0.0005)	0.8669 (0.0009)
7B	0.7174 (0.0007)	0.7489 (0.0007)	0.3907 (0.0004)	0.8321 (0.0008)
7C	0.7943 (0.0008)	0.9038 (0.0009)	0.8232 (0.0008)	0.8345 (0.0008)
8A	0.9901 (0.0010)	0.9014 (0.0009)	0.4988 (0.0005)	0.9014 (0.0009)
8B	0.9279 (0.0009)	0.8547 (0.0009)	0.4487 (0.0004)	0.9388 (0.0009)
8C	0.9628 (0.0010)	0.8887 (0.0009)	0.4311 (0.0004)	0.9715 (0.0010)
9A	0.8055 (0.0008)	0.8588 (0.0009)	0.3103 (0.0003)	0.8741 (0.0009)
9B	0.9143 (0.0009)	0.8796 (0.0009)	0.8042 (0.0008)	0.8709 (0.0009)
9C	0.9372 (0.0009)	0.9053 (0.0009)	0.4871 (0.0005)	0.9481 (0.0009)
10A	0.8168 (0.0008)	0.7865 (0.0008)	0.6180 (0.0006)	0.8659 (0.0009)
10B	0.7832 (0.0008)	0.7802 (0.0008)	0.5582 (0.0006)	0.7776 (0.0008)
10C	0.7671 (0.0008)	0.7731 (0.0008)	0.5291 (0.0005)	0.8620 (0.0009)

Table .4: Simulation Setting 11A-15C : Average Rand Index (MCSE) For Mixed Data With a Mixture of Strictly Decreasing and Periodic Functions(weighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
11A	0.9261 (0.0009)	0.9261 (0.0009)	0.5828 (0.0006)	0.9244 (0.0009)
11B	0.8644 (0.0009)	0.8960 (0.0009)	0.6277 (0.0006)	0.8392 (0.0008)
11C	0.8560 (0.0009)	0.8834 (0.0009)	0.3416 (0.0003)	0.8729 (0.0009)
12A	0.7507 (0.0008)	0.6010 (0.0006)	0.6055 (0.0006)	0.6463 (0.0006)
12B	0.6693 (0.0007)	0.5725 (0.0006)	0.5723 (0.0006)	0.5804 (0.0006)
12C	0.7044 (0.0007)	0.6354 (0.0006)	0.5659 (0.0006)	0.6002 (0.0006)
13A	0.7451 (0.0007)	0.6818 (0.0007)	0.6087 (0.0006)	0.6818 (0.0007)
13B	0.7208 (0.0007)	0.6996 (0.0007)	0.5715 (0.0006)	0.7042 (0.0007)
13C	0.7218 (0.0007)	0.7036 (0.0007)	0.6067 (0.0006)	0.7097 (0.0007)
14A	0.8947 (0.0009)	0.8830 (0.0009)	0.7206 (0.0007)	0.8477 (0.0008)
14B	0.9182 (0.0009)	0.8776 (0.0009)	0.6733 (0.0007)	0.9281 (0.0009)
14C	0.9107 (0.0009)	0.8996 (0.0009)	0.6810 (0.0007)	0.8513 (0.0009)
15A	0.8673 (0.0009)	0.7780 (0.0008)	0.6667 (0.0007)	0.7885 (0.0008)
15B	0.7299 (0.0007)	0.7404 (0.0007)	0.5564 (0.0006)	0.7525 (0.0008)
15C	0.7053 (0.0007)	0.7921 (0.0008)	0.5143 (0.0005)	0.6758 (0.0007)

Table .5: Simulation Setting 1A-10C : Average Rand Index (MCSE) For Mixed Data With Strictly Decreasing Functions-unweighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
1A	0.9806 (0.0010)	0.9806 (0.0009)	0.8739 (0.0009)	0.9806 (0.0010)
1B	0.9869 (0.0010)	0.9869 (0.0010)	0.7735 (0.0008)	0.9319 (0.0009)
1C	0.9901 (0.0010)	0.9901 (0.0010)	0.8770 (0.0009)	0.9764 (0.0010)
2A	0.9711 (0.0010)	0.9711 (0.0010)	0.8457 (0.0008)	0.9224 (0.0009)
2B	0.9618 (0.0010)	0.9368 (0.0009)	0.3810 (0.0004)	0.9141 (0.0009)
2C	0.9489 (0.0009)	0.9461 (0.0009)	0.7489 (0.0007)	1.0000 (0.0010)
3A	0.9545 (0.0010)	0.9545 (0.0010)	0.6030 (0.0006)	0.9545 (0.0010)
3B	0.9741 (0.0010)	1.0000 (0.0010)	0.4539 (0.0005)	0.9612 (0.0010)
3C	0.9887 (0.0010)	0.9887 (0.0010)	0.8764 (0.0009)	0.9887 (0.0010)
4A	0.9457 (0.0009)	0.9182 (0.0009)	0.8448 (0.0008)	0.9610 (0.0010)
4B	0.9741 (0.0010)	0.9741 (0.0010)	0.7424 (0.0007)	0.9501 (0.0010)
4C	0.9941 (0.0010)	0.9941 (0.0010)	0.8590 (0.0009)	0.9941 (0.0010)
5A	0.9901 (0.0010)	0.9901 (0.0010)	0.8689 (0.0009)	0.9901 (0.0010)
5B	0.9741 (0.0010)	0.9741 (0.0010)	0.3535 (0.0004)	0.8644 (0.0009)
5C	0.9806 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)	0.9806 (0.0010)
6A	0.8481 (0.0008)	0.7990 (0.0008)	0.5905 (0.0006)	0.8481 (0.0008)
6B	0.7287 (0.0007)	0.7287 (0.0007)	0.5069 (0.0005)	0.7305 (0.0007)
6C	0.8994 (0.0009)	0.8709 (0.0009)	0.8305 (0.0008)	0.8390 (0.0008)
7A	0.8543 (0.0009)	0.9251 (0.0009)	0.7731 (0.0008)	0.9242 (0.0009)
7B	0.8034 (0.0008)	0.8135 (0.0008)	0.6754 (0.0007)	0.8483 (0.0008)
7C	0.9497 (0.0009)	0.9477 (0.0009)	0.8087 (0.0008)	0.9582 (0.0010)
8A	0.9901 (0.0010)	0.9628 (0.0010)	0.6119 (0.0006)	0.9628 (0.0010)
8B	0.8869 (0.0009)	0.9618 (0.0010)	0.5057 (0.0005)	0.9388 (0.0009)
8C	0.9743 (0.0010)	0.9412 (0.0009)	0.7570 (0.0008)	0.8952 (0.0009)
9A	0.9804 (0.0010)	0.9533 (0.0010)	0.7966 (0.0008)	0.9537 (0.0010)
9B	0.9493 (0.0009)	0.9246 (0.0009)	0.3535 (0.0004)	0.9327 (0.0009)
9C	0.8814 (0.0009)	0.9501 (0.0010)	0.7869 (0.0008)	0.9238 (0.0009)
10A	0.8725 (0.0009)	0.8519 (0.0009)	0.6531 (0.0007)	0.8586 (0.0009)
10B	0.7816 (0.0008)	0.7364 (0.0007)	0.5236 (0.0005)	0.8099 (0.0008)
10C	0.9305 (0.0009)	0.9226 (0.0009)	0.5626 (0.0006)	0.8772 (0.0009)

Table .6: Simulation Setting 11A-15C : Average Rand Index (MCSE) For Mixed Data With Strictly Decreasing Functions-unweighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
11A	0.9614 (0.0010)	0.9261 (0.0009)	0.5440 (0.0005)	0.9323 (0.0009)
11B	0.8741 (0.0009)	0.8960 (0.0009)	0.3982 (0.0004)	0.8392 (0.0008)
11C	0.8560 (0.0009)	0.8947 (0.0009)	0.3376 (0.0003)	0.8729 (0.0009)
12A	0.7412 (0.0007)	0.7000 (0.0007)	0.6055 (0.0006)	0.7115 (0.0007)
12B	0.7014 (0.0007)	0.5725 (0.0006)	0.5723 (0.0006)	0.5804 (0.0006)
12C	0.7238 (0.0007)	0.6303 (0.0006)	0.5954 (0.0006)	0.6010 (0.0006)
13A	0.7703 (0.0008)	0.7125 (0.0007)	0.6149 (0.0006)	0.6818 (0.0007)
13B	0.7208 (0.0007)	0.6996 (0.0007)	0.5715 (0.0006)	0.7301 (0.0007)
13C	0.7158 (0.0007)	0.7687 (0.0008)	0.6067 (0.0006)	0.8390 (0.0008)
14A	0.8947 (0.0009)	0.8909 (0.0009)	0.7206 (0.0007)	0.8477 (0.0008)
14B	0.9182 (0.0009)	0.8826 (0.0009)	0.6733 (0.0007)	0.9281 (0.0009)
14C	0.9107 (0.0009)	0.9107 (0.0009)	0.6810 (0.0007)	0.8604 (0.0009)
15A	0.8885 (0.0009)	0.8139 (0.0008)	0.5618 (0.0006)	0.7885 (0.0008)
15B	0.7299 (0.0007)	0.7590 (0.0008)	0.5473 (0.0005)	0.7525 (0.0008)
15C	0.9483 (0.0009)	0.8836 (0.0009)	0.5143 (0.0005)	0.8586 (0.0009)

Table .7: Simulation Setting 1A-10C : Average Rand Index (MCSE) For Mixed Data With Strictly Decreasing Functions-weighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
1A	0.9806 (0.0010)	1.0000 (0.0010)	1.0000 (0.0010)	1.0000 (0.0010)
1B	1.0000 (0.0010)	1.0000 (0.0010)	0.7735 (0.0008)	0.9618 (0.0010)
1C	0.9861 (0.0010)	1.0000 (0.0010)	0.8770 (0.0009)	0.9764 (0.0010)
2A	0.9711 (0.9519)	0.9519 (0.0009)	0.6420 (0.0006)	0.9224 (0.0009)
2B	0.9618 (0.0010)	0.9368 (0.0009)	0.7293 (0.0007)	0.9489 (0.0009)
2C	1.0000 (0.0010)	0.9461 (0.0009)	0.7489 (0.0007)	1.0000 (0.0010)
3A	1.0000 (0.0010)	0.9545 (0.0010)	0.8366 (0.0008)	1.0000 (0.0010)
3B	0.9741 (0.0010)	1.0000 (0.0010)	0.4505 (0.0005)	0.9408 (0.0009)
3C	0.9887 (0.0010)	0.9887 (0.0010)	0.8796 (0.0009)	0.9887 (0.0010)
4A	0.9457 (0.0009)	0.9901 (0.0010)	0.8448 (0.0008)	0.9610 (0.0010)
4B	0.9741 (0.0010)	0.9741 (0.0010)	0.7543 (0.0008)	0.9501 (0.0010)
4C	0.9941 (0.0010)	0.9941 (0.0010)	0.8693 (0.0009)	0.9941 (0.0010)
5A	1.0000 (0.0010)	1.0000 (0.0010)	0.8689 (0.0009)	0.9901 (0.0010)
5B	0.9741 (0.0010)	0.9869 (0.0010)	0.7802 (0.0008)	0.9307 (0.0009)
5C	0.9806 (0.0010)	0.9901 (0.0010)	0.9901 (0.0010)	1.0000 (0.0010)
6A	0.8481 (0.0008)	0.8093 (0.0008)	0.5905 (0.0006)	0.8093 (0.0008)
6B	0.7287 (0.0007)	0.7341 (0.0007)	0.5275 (0.0005)	0.7305 (0.0007)
6C	0.7630 (0.0008)	0.7630 (0.0008)	0.6487 (0.0006)	0.8390 (0.0008)
7A	0.8543 (0.0009)	0.9251 (0.0009)	0.7685 (0.0008)	0.9242 (0.0009)
7B	0.8034 (0.0008)	0.8135 (0.0008)	0.6754 (0.0007)	0.8483 (0.0008)
7C	0.9497 (0.0009)	0.9006 (0.0009)	0.8691 (0.0009)	0.9582 (0.0010)
8A	0.9901 (0.0010)	0.9143 (0.0009)	0.5816 (0.0006)	0.9901 (0.0010)
8B	0.8869 (0.0009)	0.9618 (0.0010)	0.5057 (0.0005)	0.9388 (0.0009)
8C	0.9743 (0.0010)	0.9412 (0.0009)	0.7570 (0.0008)	0.8952 (0.0009)
9A	0.9804 (0.0010)	0.9430 (0.0009)	0.7242 (0.0007)	0.9537 (0.0010)
9B	0.9493 (0.0009)	0.9246 (0.0009)	0.3535 (0.0004)	0.9327 (0.0009)
9C	0.9529 (0.0010)	0.9501 (0.0010)	0.8075 (0.0008)	0.9057 (0.0009)
10A	0.8725 (0.0009)	0.8519 (0.0009)	0.4931 (0.0005)	0.8586 (0.0009)
10B	0.7816 (0.0008)	0.7364 (0.0007)	0.5236 (0.0005)	0.8099 (0.0008)
10C	0.9305 (0.0009)	0.9226 (0.0009)	0.6055 (0.0006)	0.8772 (0.0009)

Table .8: Simulation Setting 11A-15C : Average Rand Index (MCSE) For Mixed Data With Strictly Decreasing Functions-weighted).

Simulation Number	Complete (MCSE)	Average (MCSE)	Single (MCSE)	Ward (MCSE)
11A	0.9446 (0.0009)	0.9261 (0.0009)	0.5440 (0.0005)	0.9323 (0.0009)
11B	0.8741 (0.0009)	0.8960 (0.0009)	0.3982 (0.0004)	0.8392 (0.0009)
11C	0.8560 (0.0009)	0.8947 (0.0009)	0.3376 (0.0003)	0.8729 (0.0009)
12A	0.7412 (0.0007)	0.7000 (0.0007)	0.6055 (0.0006)	0.7115 (0.0007)
12B	0.7263 (0.0007)	0.5725 (0.0006)	0.5723 (0.0006)	0.5804 (0.0006)
12C	0.7238 (0.0007)	0.6303 (0.0006)	0.5954 (0.0006)	0.6010 (0.0006)
13A	0.7703 (0.0008)	0.6956 (0.0007)	0.6160 (0.0006)	0.6818 (0.0007)
13B	0.7208 (0.0007)	0.6996 (0.0007)	0.5715 (0.0006)	0.7301 (0.0007)
13C	0.7158 (0.0007)	0.7687 (0.0008)	0.6067 (0.0006)	0.7133 (0.0007)
14A	0.8947 (0.0009)	0.8909 (0.0009)	0.7206 (0.0007)	0.8477 (0.0008)
14B	0.9182 (0.0009)	0.8826 (0.0009)	0.6733 (0.0007)	0.9281 (0.0009)
14C	0.9107 (0.0009)	0.9107 (0.0009)	0.6810 (0.0007)	0.8604 (0.0009)
15A	0.8885 (0.0009)	0.8139 (0.0008)	0.5424 (0.0005)	0.7885 (0.0008)
15B	0.7299 (0.0007)	0.7590 (0.0008)	0.5473 (0.0005)	0.7525 (0.0008)
15C	0.9483 (0.0009)	0.8836 (0.0009)	0.5143 (0.0005)	0.8586 (0.0009)

VITA

OBED KOOMSON

Education:	M.S. Mathematical Sciences (Statistics), East Tennessee State University Johnson City, Tennessee 2018 B.Sc Actuarial Science, Kwame Nkrumah University Science and Technology, Kumasi, Ghana,2008– 2012.
Professional Experience:	Graduate Teaching Assistant, East Tennessee State University Johnson City, Tennessee, 2017–2018. Operations Manager, Medi-Life Co. Ltd. Techiman, Ghana, 2015–2016. Store Supervisor, Kasapa Health Services. Techiman, Ghana, 2013–2015. Mathematics Teacher, Techiman Senior High School. Techiman, Ghana, 2012–2013.
Professional Development:	Statistical and Mathematical, Software: SAS, R, SPSS, Minitab, Python. Microsoft Office Suite: MS Access, VBA, Word, Excel, PowerPoint, Outlook.