8-2018

Clustering Mixed Data: An Extension of the Gower Coefficient with Weighted L2 Distance

Augustine Oppong
East Tennessee State University

Clustering Mixed Data: An Extension of the Gower Coefficient with Weighted L_2 Distance

---

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

Augustine Oppong

August 2018

JeanMarie Hendrickson, Ph.D., Chair

Robert Price, Ph.D.,

Nicole Lewis , Ph.D.

Keywords: Clustering, Mixed data, Extended Gower coefficient

ABSTRACT

Clustering Mixed Data: An Extension of the Gower Coefficient with Weighted L_2 Distance

by

Augustine Oppong

Sorting out data into partitions is increasing becoming complex as the constituents of data is growing outward everyday. Mixed data comprises continuous, categorical, directional functional and other types of variables. Clustering mixed data is based on special dissimilarities of the variables. Some data types may influence the clustering solution. Assigning appropriate weight to the functional data may improve the performance of the clustering algorithm. In this paper we use the extension of the Gower coefficient with judciously chosen weight for the L_2 to cluster mixed data.The benefits of weighting are demonstrated both in in applications to the Buoy data set as well simulation studies. Our studies show that clustering algorithms with application of proper weight give superior recovery level when a set of data with mixed continuous, categorical directional and functional attributes is clustered. We discuss open problems for future research in clustering mixed data.

ACKNOWLEDGMENTS

First and foremost, I would like to thank the Almighty God for how far He has brought me and seeing me through this thesis. I would like to acknowledge Dr. JeanMarie Hendrickson for her guidance and support throughout this project. At some point it got tough, it was her encouragement and productive feedback that kept me going. I am indebted to her for accepting to be my adviser. I would also like to thank Dr.Robert Price and Dr.Nicole Lewis for agreeing to be of my committee. Finally, I would like to acknowledge my family, friends and loved ones for their immense support, love and prayers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Large amount of data are collected every year. The Havard Business School in their article "From data to Action" emphasized that we are confronted with the problem of identifying what constitutes data. The boundaries we have today with regards to what we can digitize and analyze are growing outward every day [7]. Data exist in different forms and this has led to the intensive study of clustering. Clustering has been marked as a basic method of data mining for the unearthing of valuable knowledge. Other fields of study have successfully exploited its benefit. We see its success in pattern recognition [3] biology, psychology, psychiatry, archaeology, geology, geography, marketing, image processing and information retrieval [4]. In recent times clustering has been applied in proteomic studies where a mass spectrometer records macromolecular observation values continuously across a domain of mass/charge ratio values. The problem of clustering may generally be stated as follows:

Given a set of raw data points, sort them into a set of classes such that the classes are similar as possible [1].

In this sense clustering provides a summarized and compact form of data to benefit the ordinary user and the researcher as well. Clustering has proved essential to address "big data". The concept of mixed data sets emerge when variables under consideration consist of several types, e.g., continuous, cat-

egorical, functional, directional, etc. This heterogeneity of data is driving research into mixed data clustering.

1.1 Background on Cluster Analysis

We define clustering as a way of grouping data such that objects in the same group look similar and objects in different groups are heterogeneous, according to some standard metric. A series of work focusing on existing clustering algorithms has been done. Some of this work provide comparisons, both from a theoretical and an empirical outlook, on the performance of the individual algorithm. The potency of the candidate algorithm is accessed through a number of internal and external validity metrics, stability, runtime and scalability tests. In general, existing clustering algorithms are categorized broadly as follows: Partitioning based, Hierarchical-based, density-based, Grid-based and Model-based. The partition based includes the K-means, K-medoids k-modes etc. [29].

In K-means, we consider the center as the average of all points. This algorithm partitions data into k groups by minimizing some criterion; the within-group sum of squares over all variables is often used as the minimizing criterion [6]. It starts by selecting some K points as initial centroids. Each point is then assigned to the nearest centroid depending on some chosen proximity measure. This forms a cluster and the centroids for each cluster are updated. The algorithm repeats these steps until a stopping criterion is

reached. The challenges coupled with the K -means include being sensitive to outliers. Another challenge is that it works only when the mean of a cluster is specified. Also the number of groups must be specified in advance. As a result others methods such as the K-medoids and K-modes could be alternatives. Cuesta-Albertos et al (1997) proposed trimmed k-means clustering. This alternative method calculates the cluster means based on only $N(1-\alpha)$ observations out of the entire N observations where $0 < \alpha < 1$ [8].

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. This method can either be divisive or agglomerative. The divisive approach begins with the entire data set as one cluster and repeatedly splits the data into the most appropriate cluster. Agglomerative clustering on the other hand starts with one object in each cluster and repeatedly merges the data into two or more appropriate clusters. The dissimilarities between objects are usually measured by some distance. One of the commonly used distance is the Euclidean distance. We define the Euclidean distance between two p-dimensional observation $\mathbf{x}$ and $\mathbf{y}$ as

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

Other distance measures include the Minkowski distance of order p, defined as

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

, and the Canberra metric

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

. There are methods or criteria applied in hierarchical clustering to determine the distance between pairs of observations as a function of the pairwise distances. Among them are single linkage, complete linkage average linkage and the ward's method. For two clusters A and B, single linkage defines the distance between A and B as

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

. Complete linkage clustering defines distance between A and B as

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

. Average linkage clustering defines 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 while d_{ij} is the distance between objects. There is another method known as the Ward's method. In Ward's approach, the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. This criterion seeks to minimize the total within-cluster variance. To implement this method, at each step find the pair of clusters

that leads to minimum increase in total within-cluster variance after merging [12]. The hierarchical method has a major setback. Once an object merges or splits into a cluster, it cannot be reversed.

The density based methods have objects separated based on their area or region of density, congruence and perimeter. Clusters develop and move in the direction that density leads to. Density-based algorithms are capable of discovering clusters of arbitrary shapes. This method provides a natural protection against outlying observations. Here, to put an object into a cluster, its overall density is evaluated to determine the functions of datasets that influence this particular data object. Some algorithms such as DBSCAN, OPTICS, DBCLASD and DENCLUE make use of this approach to discover clutsers of arbitrary shape [29].

Under the grid-based methods, the entire space of observations is partitioned into a grid. This method has been noted for its fast processing time because it goes through the dataset once to calculate the statistical values for the grids. The accumulated grid-data make these clustering procedures independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid. So the clustering is not performed on the database directly. The performance of a grid-based method depends on the size of the grid, which is usually much less than the size of the database. This method however does not perform

well on highly irregular data. Some examples of algorithms that use this approach include Wave-Cluster and STING [29]

The model based clustering is based on the assumption that the data is generated by a mixture of underlying probability distributions. It encompasses analyses of finite mixture densities [6]. This method models the set of clusters and find the best fit of the data to the model. We write the mixture distribution or the probability density function of x_n as

$$p(x_n) = \sum_{i=1}^{K} \pi_i p(x; \theta)$$

where $\pi_1 ... \pi_K$ are the mixing probabilities, $p(x; \theta)$ is the component distribution of parameter, θ. The mixing probabilities must satisfy

$$0 \leq \pi_K \leq 1 \; and \; \sum_{i=1}^{K} \pi_K = 1$$

.

The model based clustering reduces clustering to estimation of model parameters. These parameters are estimated using expectation- maximization algorithm or other Bayesian estimation methods. Of course there are other model-based procedures. Banfield and Raftery [25] extended the classification maximum likelihood procedure which was originally proposed by Scott and Symons [31]. The model based approach to clustering arose to address some of the challenges of the K-means and agglomerative hierarchical clustering; it must be emphasized that the the former allows possible inference

whiles the latter serve as exploratory tools.

Most clustering algorithms work with numeric data but work has been done to include categorical data [23]. With the ever growing constituents of data, mixed data containing both numeric and categorical characters exist. The traditional approach is to transform the categorical data into numeric values and proceed with clustering based on the existing numerical algorithm. Also it could be done by clustering the categorical data directly where we assign a distance of 1 if the two values are distinct and a distance of 0 if the two values are identical. The setback here is that the clustering does not reveal the true similarity structure of the data set since the two methods fail to take into account the similarity information contained in between categorical values. Hsu, 2006; Hsu and Wang, 2005 made similar assertions [33].

Other types of data arise as curves or functions of one or several independent variables. These data types are normally referred to as functional data. Clustering of such data may depend on some characteristics of curves such as positions, shapes and derivatives. For a variable to be considered functional we must be able to determine the quantity existing at any time (or other charting variable) along the interval T. Most clustering algorithms depend on some dissimilarity measure between objects to guide the algorithm. When it comes to functional data, the ideal distance measure between two curves,

$y_i(t)$ and $y_j(t)$ measured on some domain, [0,T] is the squared L_2 between the two curves:

$$d(i,j) = \int_0^T [y_i(t) - y_j(t)]^2 dt.$$

In dealing with functional data, each discrete datum is converted to a continuous functional observation by a method of smoothing. In Ferreira et.al (2009), they applied the B-spline to smooth each functional observation. The goal was to remove noise and not change drastically the structure of the curves before clustering. The L_2 distance between two curves was approximated by the trapezoidal rule. They argued that the data in practice consist of discrete values representing measurements along continuous curves. We have the function h(t) along the domain [0,T] approximated as

$$I_n = \frac{T-0}{2n}[h(0) + 2h(t_1) + ... + 2h(t_n) + h(T)],$$

where n is the number of measurement points used in the approximation.[12]

2 REVIEW OF PREVIOUS METHODOLOGY FOR CLUSTERING MIXED DATA

Several methods have been proposed to cluster mixed data. The available options include doing a separate clustering on each variable type. There are setbacks to this approach as there cannot be common ground for these clusters.There is no concrete agreement from the conclusions made from the clusters. Other methods involve converting all variables into a solitary variable and then proceed with the cluster analysis. The disadvantage with this method is that a great deal of information is lost and the true clustering is affected [16].

Cluster analysis usually focuses on finding the dissimilarity between objects, but we can reach the same conclusion also by focusing on the similarity between objects. J.C Gower (1971) proposed a coefficient to measure the similarity between two objects based on mixed data [17]. Several works have been done since then on the use of this coefficient. In 2006, Chae, Kim and Yang assigned weights to the variable types to solve the problem which existed in Gower's proposed formula. They note that assigning weights to either one of the variable types leads to clustering with dominance of one variable type. Appropriate weights were allotted based on the characteristics of the data under consideration. This was done in order to protect or favor the phenomenon of dominance of one variable type. The dissimilarity

measure they define is stated as follows

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

$$= \tau_{ij} \sum_{l=1}^{c} \frac{1}{c}(\frac{|x_{il} - x_{jl}|}{R_l}) + (1 - \tau_{ij})\sqrt{(1 - \frac{\sum_{l=c+1}^{r} s_{ijl}}{\sum_{l=c+1}^{r} w_{ijl}}}$$

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

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

where $-1.0 \leq \rho_{ij}^{c}$ 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, for variables provided the comparison between the ith and jth objects is valid for the lth variable. They acknowledged that they employed the Pearson correlation coefficient and the product moment correlation coefficient. The Pearson correlation coefficient is substitute for ρ_{ij}^{c} and the product moment correlation is used for ρ_{ij}^{d} [18].

A method that came to be known as a generalized Minkowski metric is capable of handling continuous, discrete, ordinal and nominal variables[15]. The method has proved to work with tree structured variables with a finite

set of nominal values. It turns out that the proposed dissimilarity measure is a metric distance. Ichino and Yaguchi (1994) recommend alternative kind of the measure may that be normalized to incorporate variables with different measurement scales or the application of different weight measures for the different variable types .

Friedman and Meulman (2004) elaborate two algorithms for clustering objects on subsets of attributes (COSA). This method was derived from already known or existing distance measures. In their work they state that the COSA algorithm focuses primarily on clustering objects based on uncertain similar joint values for the attributes. They thought through such a method for potential setbacks and came up with some modifications. The minimal modification to the COSA algorithm permitted it to be used for clustering objects based on certain values of the attributes. This they refer to as "Single-target clustering". Likewise, it can be used to cluster objects based on two extreme values which they refer to as "Dual-target clustering" [10].

The model-based method seeks to optimize the fit between the data and some mathematical model. The underlying assumption is that data is generated by some underlying probability distribution. This provides a means of automatically ascertaining the number of clusters based on standard statis-

tics. Everitt (1988) proposed a clustering model which obtained credit as one of the earliest model based approach for clustering mixed data. The model could work for data containing both continuous and ordinal or nominal variables. Several assumptions were made regarding this model. It is assumed that an observed vector $\mathbf{x}$, consisting of $p+q$ random variables assumed the 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 multivariate normal density. Everitt continued that the binary and ordinal variables can be thought of as coming from underlying continuous distributions. The q categorical or ordinal variables are obtained by setting certain threshold values as cut-off points.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,

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

Where the $\delta_{ijl}, i = 1, ..., k, j = 1, ..., q, l = 2, ..., t_j$ are the threshold values used to construct the ordinal variables, $z_1, z_2, z_3, ..., z_q$ from the continuous variables $x_{p+1}, x_{p+2}, x_{p+3}, ..., x_{p+q}$ Everitt(1988) broached and defined the

density function of the form

$$g(x,z) = \sum_{i=1}^{k} p_i MVN_{(p)}(\mu_i^{(p)}, \Sigma) \int_{a_1}^{b_1} ... \int_{a_q}^{b_q} MVN_q(\mu_i^{(q|p)}, \Sigma_{(q|p)}) dy_1, ...dy_q$$

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 is 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 covarinace matrix of $x_1, ..., x_p$; Σ_q is the covariance matrix of $x_{p+1}, ..., x_{p+q}$. The problem of model based cluster analysis boils down to the estimation of parameters of density function. There is the need to estimate the parameters for a given set of observations in order to determine the probabilities for assigning appropriate clusters to the observation. In order to estimate the parameters we maximize the log-likelihood function

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

There are other approaches to mixture model clustering which are extensions of previously known methods. The observations should be in the form of an $n \times p$ matrix. The observations here come from variables which are a random sample of the form $f(x) = \Sigma \pi_k f_k(x)$ [19]. A later model called the MULTIMIX model for the ith observation is given as

$$f(x_i; \phi) = \Sigma_k^K \pi_k \prod_{l=1}^{L} f_{kl}(\bar{x}_{il}; \theta_{kl}$$

where θ_{kl} contain the parameters of f_{kl} and we treat π_k as mixing probabilities. If the distribution of f_{kl} is known then there could be a known algorithm for estimating the parameters. For instance if the f_{kl} belongs to the exponential family, the model's parameters can be estimated using the Expectation Maximization (EM) algorithm.[19]

2.1 Extended Gower Coefficient

Hendrickson and Hitchcock created an extension of the Gower coeffient to cluster mixed data types. 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. 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 directional, then

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

where θ_i is the angle measured on object i [21] If the variable is interval scaled, then

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

2.2 Proposed Work

It is noted that, with regards to actual work done on clustering in literature, hierarchical clustering is the most widely used clustering method in practice [11]. We are going to apply hierarchical clustering algorithms with the Extended Gower coefficient to cluster mixed data. We will examine both the effects of unweighted and weighted dissimilarities to the functional data. We want to know how functional data affect our clustering results when combined with other variables. In particular we will employ the inverse-variance weight used by Chen et. al. We will also make use of another weight function called the CV-optimal weight in our simulation studies. This weight function was proposed by Huaihou et. al and it seeks to minimize the coefficient of variation of a random vector θ [20]. We will describe the weight in later chapters. Various simulation studies will be done as well as applications to real data set.

3 APPLICATION TO REAL DATA

We now illustrate and present the extension of the Gower coefficient with the appropriate weight on a real data set. The data we are using are obtained from the National Buoy Data Center (NBDC) historical data page http://www.ndbc.noaa.gov at the National Oceanic and Atmospheric Administration (NOAA) web site. The data has many variables but the variables we will consider are wind direction, wind speed, air temperature, and water temperature. We also include the latitude, longitude and water depth that corresponds to each buoy. Since some of the buoys contain too many missing values we delete such buoys. If the degree of missing values can be tolerated ,such as missing values in variables that were time series, we took the average of the observations before and after the missing observation.The number of observations and the measurement times for the functional data varied from buoy to buoy. We made some adjustment to the time index to be the number of minutes since January 1, 2011 at 12:50 AM. The time in the original data sets was given as year, month, day, hour, and minute We also considered the time zone (a nominal variable) for each buoy. The time zones in our buoy data were Central, Eastern, and Pacific. To make allowance for this variable in the Gower coefficient, we ascribe a numerical label for each time zone, 'Eastern'=1, 'Central'=2, and 'Pacific'=3. We would treat wind direction, wind speed, air temperature, and water temperature as functional

variables. We also employed the B-spline smoother to introduce some degree of smoothness to the curves since each buoy had different measurement points We would also treat latitude and longitude as directional variables. To transform from Cartesian coordinates to polar coordinates, we make the following transformations; $x = rcos\theta$ and $y = rsin\theta$ where $r = \sqrt{x^2 + y^2}$ .Since latitude and longitude are measured using the intersection of the prime meridian and the equator as the reference point, and since all our data lie in the same region, we use 0° latitude, 0° longitude as our reference point. We can solve for θ by using the formula $\theta = arccos\frac{x-x_0}{\sqrt{(x-x_0)^2+(y-y_0)^2}}$, where x is the latitude, y is the longitude for each buoy, x_0 is the latitude, y_0 is the longitude at the reference point. To illustrate this with an example we choose buoy 51000 which is located at $23.546°N154.056°W$. We find the value of θ for this buoy as; $\theta = arccos\frac{154.056-0}{\sqrt{(23.546-0)^2+(154.056-0)^2}} = 0.152$ radians.

3.1 B-Splines

B-spline functions comprise flexible bands that pass through a number of points that are called control points and create smooth curves. By these functions we are able to develop and manage complex surfaces through a number of knots. The spline functions are continuous at the knots. In this way we construct curves from a given set of points by making the curve pass through the points. Curves consisting of just one polynomial or rational

segment has some shortcomings. For instance a high degree is required in order to satisfy a large number of constraints e.g., (n 1)-degree is needed to pass a polynomial Bezier curve through n data points. However, high degree curves are inefficient to process and are numerically unstable; Also a high degree is required to accurately fit some complex shapes. A way to address some of these setbacks is to make use of curves that are piecewise polynomial [27]. Our aim is to achieve some level of smoothness by joining together neighboring curves or line segments. We use the bs() function in R to produces B-splines. The curves derived from the functional observations were obtained this way. See Figures 3.1 and 3.2. This proves to be a computationally efficient way to compute cubic regression splines.

3.2 The Rand Index and Adjusted Rand Index

We assessed the clustering solution using the Adjusted rand and the Rand indices. The rand index is a commonly used clustering accuracy metric introduced by Rand (1971). The rand index takes into account the number of instances that exist in the same cluster and in different clusters between two cluster solutions. The rand index is defined as follows

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

where n_{11} is the number of pairs of instances clusters that are placed in the same cluster in both, n_{00} is the number of pairs of instances that are different

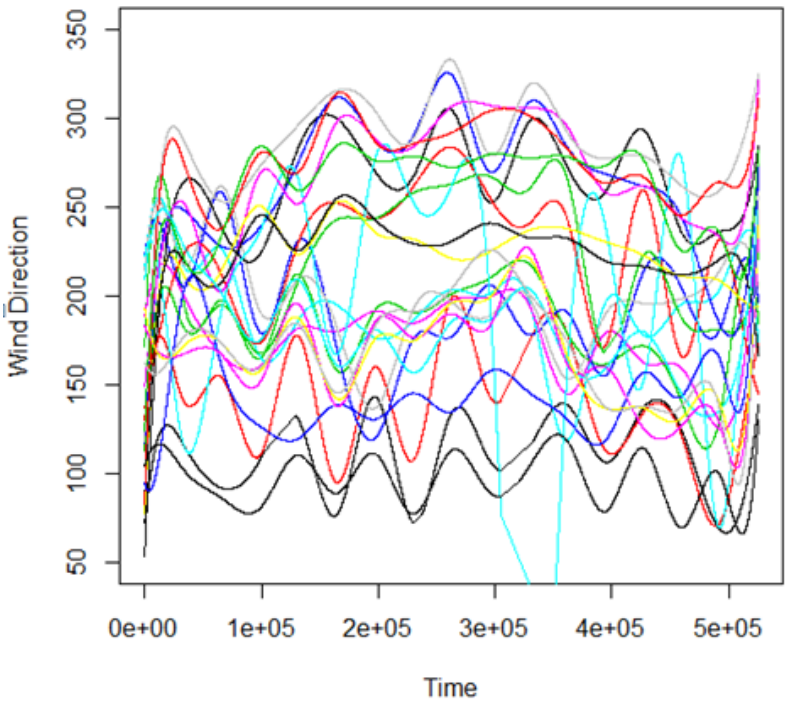

Figure 3.1: Curves showing Wind direction

clusters, n_{10} is the number of pairs of instances that are in the same cluster in A, but in different clusters in B, n_{01} is the number of pairs of instances that are in different clusters in A, but in the same clusters in B. A is one clustering result from the data and B is another clustering result from the same data. We have $0 \leq R \leq 1$. Values close to 0 indicates that the two data clusterings do not agree. A value close to 1 indicates that data is clustered in nearly the same way, with a value of 1 indicating the two cluster solutions are the same. One setback of the Rand Index is that the expected value of the of the Rand Index between two random clusters may be different. This setback would be

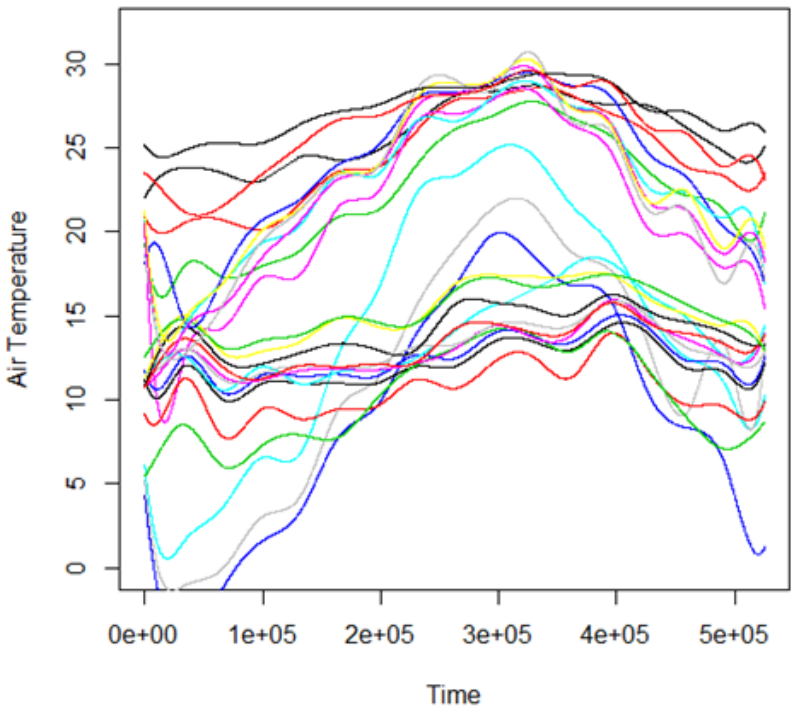

Figure 3.2: Curves showing Air temperature

addressed with the Adjusted Rand Index (ARI). The Adjusted Rand index is another validation measure used to compare the clustering solution of two clusterings. Here, an account is taken of randomness or chance of overlap in the clustering. The Adjusted Rand Index (ARI) is given as

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{N}{2}},$$

where n_{ij} is the number of objects that are common to clusters A_i and B_j and A_i is the i^{th} cluster in the first clustering and B_j is the j^{th} cluster in the second

clustering, $a_,$ and $b_j s$ are marginal sums, $a_i = \sum_j n_{ij}$ and $b_j = \sum_i n_{ij}$ and $\sum_{ij} n_{ij} = N$. [26]. We note that although the Rand index is always positive, the Adjusted Rand Index can produce negative values if the expected value is greater than the index

3.3 Weighted L_2 in the Extended Gower coefficient

The Extended Gower coefficient permits the inclusion of the directional and functional variables with the other types of variables. We defined and used the following dissimilarity for each variable type. We used the L_1 for the continuous or the interval-scaled variable . As defined previously the L_1 distance between observations x_i and x_j for the f^{th} variable is

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

. The dissimilarity measure for the functional variable is the L_2 distance,

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

We consider the weighted L_2 distance which shall be employed to measure distances for functional data;

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

. In R we use the function metric.lp in the fda.usc. package. This function calculates an approximate L_p distance for functional data using Simpson's

rule [28]. The function fdata was also employed to transform the fitted values from fitting the B-splines for each functional variable to a functional data object. For our functional variables, we calculated the variances from each of the observed curves and used it to compute the inverse variance weight. We used the Ackerman distance to calculate the dissimilarity measure for the directional variable. If θ_i is the angle measured on object i, we have

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

If the variable is binary or nominal we use,

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

The dissimilarity for the combined set of variables is calculated as follows; the distance between objects i and j is

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

where $\delta_{ij}^{(f)}$ is 1 if both measurements x_{if} and x_{jf} for the fth variable are non-missing, and 0 otherwise (Kaufman Rousseeuw, 1990). So if we have that both measurements x_{if} and x_{jf} for the fth variable are non-missing the dissimilarity between the ith and jth objects will simply be the sum of all of the dissimilarities calculated for the ith and jth objects, divided by the sum of the number of variables. With our variables, we created a 26 X 26 dissimilarity matrix containing the dissimilarity measures defined above. We

then used the hclust package in R on this matrix to perform hierarchical clustering on the matrix. Since there are seven regions we make a choice of 7 clusters.

3.4 Summary of Results

When equal weight was applied to the functional data the Adjusted Rand Index is 0.2604 and the Rand Index is 0.8185. The application of the inverse variance weight to the functional variable saw an increase of 29.11% in the Adjusted Rand Index from 0.2604 to 0.3332. The Rand index increased by 3.01%. from 0.8185 to 0.8431. The inverse variance weighted functional data outperformed the equal weight functional data. Table 3.1 shows the clustering solution of the weighted data.

Table 3.1: Clustering solution with Extended Gower coefficient.

Station	Region	Cluster label
41046	Western Atlantic	1
41047	Western Atlantic	1
41078	Western Atlantic	2
44007	Northeast USA	3
44009	Northeast USA	3
41004	Southeast USA	2
41012	Southeast USA	2
44020	Southeast USA	3
46012	Northeast USA	4
46027	Northeast USA	5
46041	Northeast USA	5
46042	Northeast USA	4
46059	Northeast USA	6
46011	Southeast USA	4
46025	Southeast USA	5
46028	Southeast USA	4
46053	Southeast USA	5
46054	Southeast USA	4
46086	Southeast USA	5
42020	Florida/Gulf of Mexico	1
41012	Florida/Gulf of Mexico	2
42036	Florida/Gulf of Mexico	2
42039	Florida/Gulf of Mexico	2
42040	Florida/Gulf of Mexico	2
42056	Western Caribbean	1
42055	Western Caribbean	7

As can be seen from Table 3.1, the 7 cluster solution produced some correct clusters as well as some misclassified clusters. The buoys were in regions Western Atlantic, Northeast USA, Northwest USA, Southwest USA

and Western Caribbean. All stations which are located in Gulf of Mexico were misclassified into Northeast USA. Stations 44007 and 44009 which are in Southeast USA were placed in Northeast USA region. The table shows the clustering solution of the weighted data which is similar to the equally weighted data. But we observed from the dendrograms that buoy number 26 which geographically is located along the West coast was clustered by itself in the original data (See Figure 3.3). But with the weighted data it moved and aligned itself with other buoys along the west coast. This improved cluster solution indicates the benefit of the weighted data (See Figure 3.4).

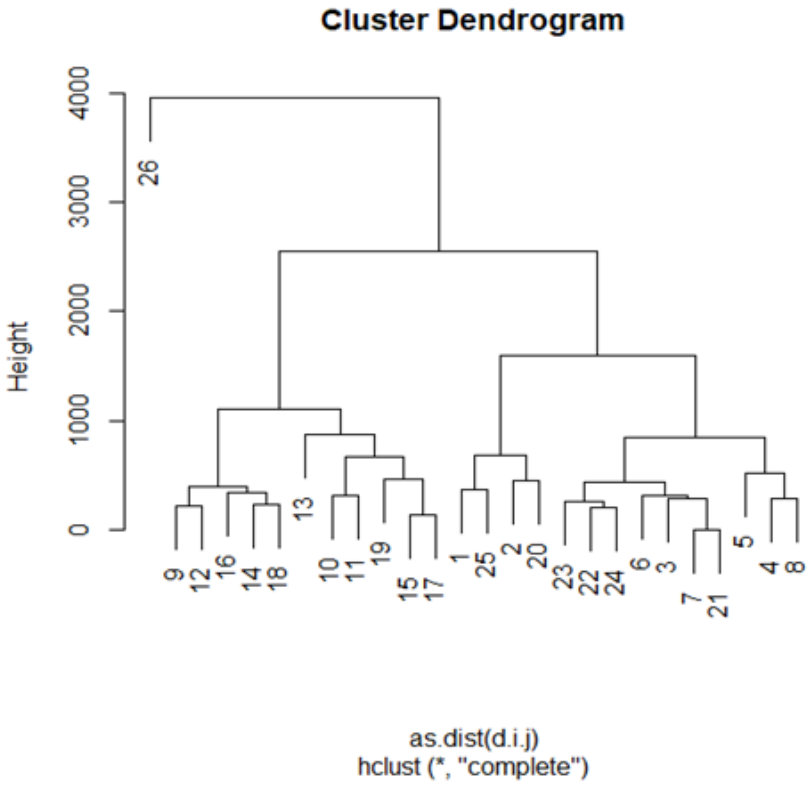

Figure 3.3: A dendrogram showing extension of the Gower coefficient

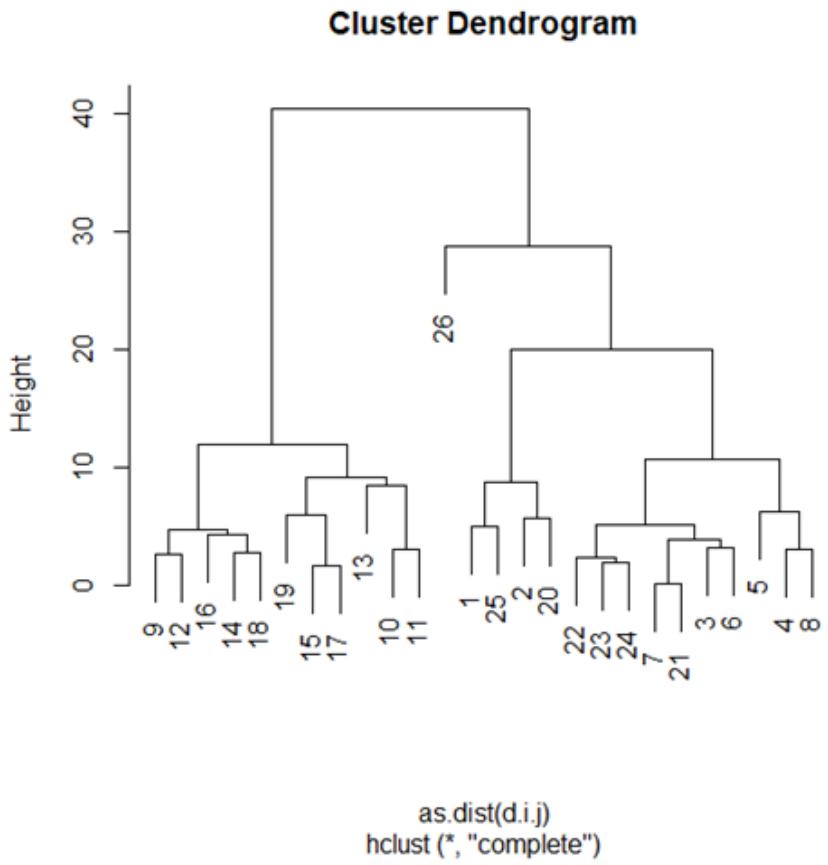

Figure 3.4: A dendrogram showing the clustering solution of the weighted data

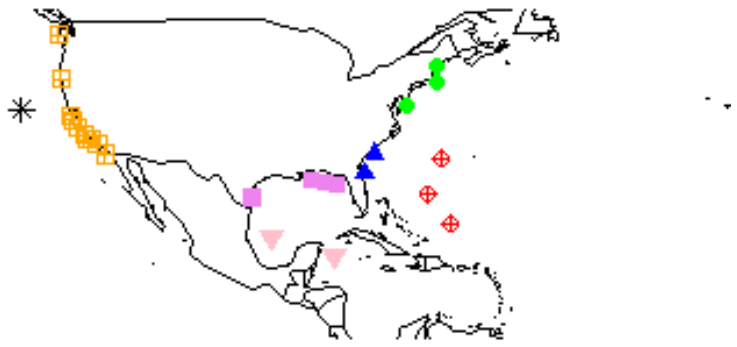

Figure 3.5: A US map showing the location of the buoys

A US map shows the location of the buoys in Figure 3.5. It can be realized that the clustering solution classified most of the buoys into regions relatively close geographically to their "true regions". Therefore the 7-cluster solution does not deviate much from the reality. The clustering solution does a satisfactory job.

4 SIMULATION STUDY

We conducted simulation studies to compare the clustering results of mixed data with two different weight functions. The data generated consists of continuous, categorical, directional and functional observations. The extended Gower coefficient allows this data set to be clustered concurrently. This data simulated is similar to that of Hendrickson and Hitchcock (2012). The categorical data was generated using the *sample.int* function in R. This function enables us to sample from multinomial probability function with replacement from five categories and has 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 $N = \sum_1^5 x_i, p_i$ is the probability for each category [13]. The cluster sizes were made to differ and for each cluster we made use of different probability vectors, in order to simulate clusters of data that have different probabilities of coming from each category. In this simulation study,the probability vectors were (0.8,0.05,0.05,0.05,0.05) where we have one principal or dominant category and we chose the probability vectors (0.2,0.2,0.2,0.2,0.2) to represent equally likely categories. The continuous variable was simulated from a normal distribution with mean μ and standard deviation σ. For the purpose of the study we chose the following values for μ and σ: σ was fixed at 100 for all four clusters , $\mu = 5000$ and for cluster 1, for clus-

ter 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. Here for instance $k = 5, k = 20$ and $k = 50$. As the value of k increases, the separation between clusters becomes greater. The directional variable θ, is 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 [22]. 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

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

[14]. In R, the rvonmises function in the circular package is used to simulate this distribution. For the 4 different clusters, we picked the following values 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 $k = 2.5$. There are three different groups of signal curves that could be used for generating functional

data. These signal curves were the same signal curves used by Hitchcock and Ferreira (2009). They were chosen according to their description to lie within a reasonable proximity to each other to make the clustering solution good as possible. As presented by Hitchcock and Ferreira (2009), the first group is made up of some form of periodic data. The first group is defined as follows:

$$\mu_1(t) = \frac{1}{28}(t) + exp(-t) + \frac{1}{5}sin(t/3) + 0.5$$
$$\mu_2(t) = \frac{1}{20}(t) + exp(-t) + \frac{1}{5}sin(t/2)$$
$$\mu_3(t) = \frac{1}{15}(t) + exp(-t) + \frac{1}{5}cos(t/2) - 1$$
$$\mu_4(t) = \frac{1}{18}(t) + exp(-t) + \frac{1}{5}cos(t/2)$$

The second group of signal curves had no periodic tendencies and strictly decreasing defined as follows:

$$\mu_1(t) = 50 - (t^2/500) - 7ln(t)$$
$$\mu_2(t) = 50 - (t^2/500) - 5ln(t)$$
$$\mu_3(t) = 50 - (t^2/750) - 7ln(t)$$
$$\mu_4(t) = 50 - (t^2/250) - 2ln(t)$$

The third group had a decreasing trend and contained a mixture of some periodic tendencies and strictly decreasing functions. They are defined as

follows:

$$\mu_1(t) = -t/2 + 2sin(t/5)$$

$$\mu_2(t) = -t/2 + 2cos(t/3)$$

$$\mu_3(t) = -t^2/250 - 4ln(t)$$

$$\mu_3(t) = -t^2/250 - 2ln(t)$$

The three groups of the signal curves are plotted in Figures 6, 7 and 8 respectively

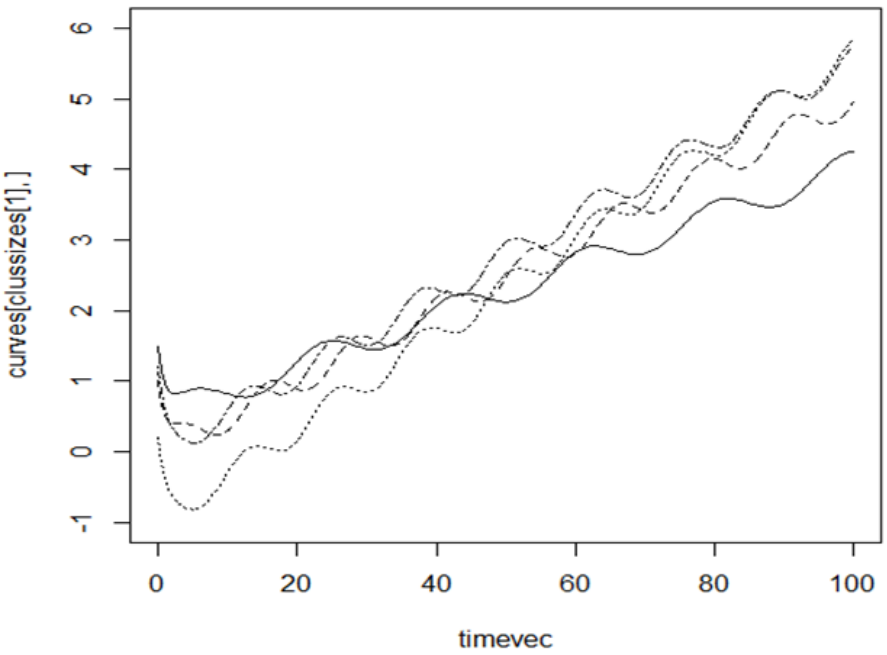

Figure 4.1: Group 1 signal curves.

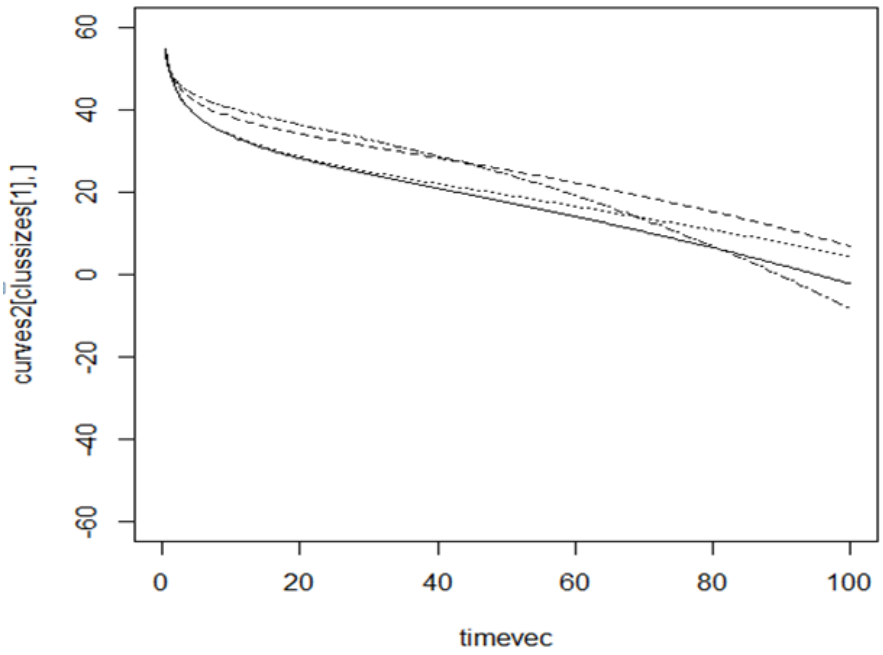

Figure 4.2: Group 2 Signal curves

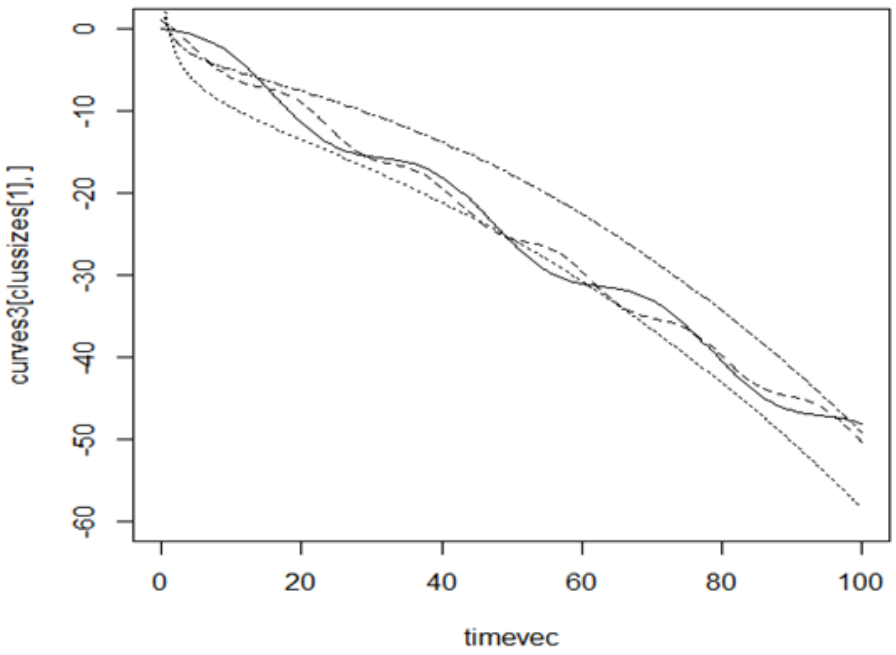

Figure 4.3: Group 3 Signal curves

Following the same framework provided by Hitchcock and Ferreira (2009) and the work of Hitchcock and Hendrickson, we generated 30 discretized curves based on the 4 signal functions from the groups above. The data was simulated over 200 points from t = 0 to t = 100 in increments of 0.5 except for the clusters containing the ln(t) in which case the data was simulated over 201 points from t = 0.5 to t = 100 by increments of 0.5. We introduced a random error term to the signal functions thereby making allowance for some variation within each cluster. A discretized approximation of the stationary Ornstein-Uhlenbeck process was used in this case. This process is a Gaussian

process with mean zero and the covariance between the errors measured at points t_m and t_n is $\sigma^2(2\beta)^{-1}exp(-\beta|t_m - t_n|)$ (Ferreira and Hitchcock, 2009). We kept the drift variable β, at 0.5 and let $\sigma = 1.75$ for small distance between the clusters and $\sigma = 1$ for large distance between the clusters. See Figure 4.4

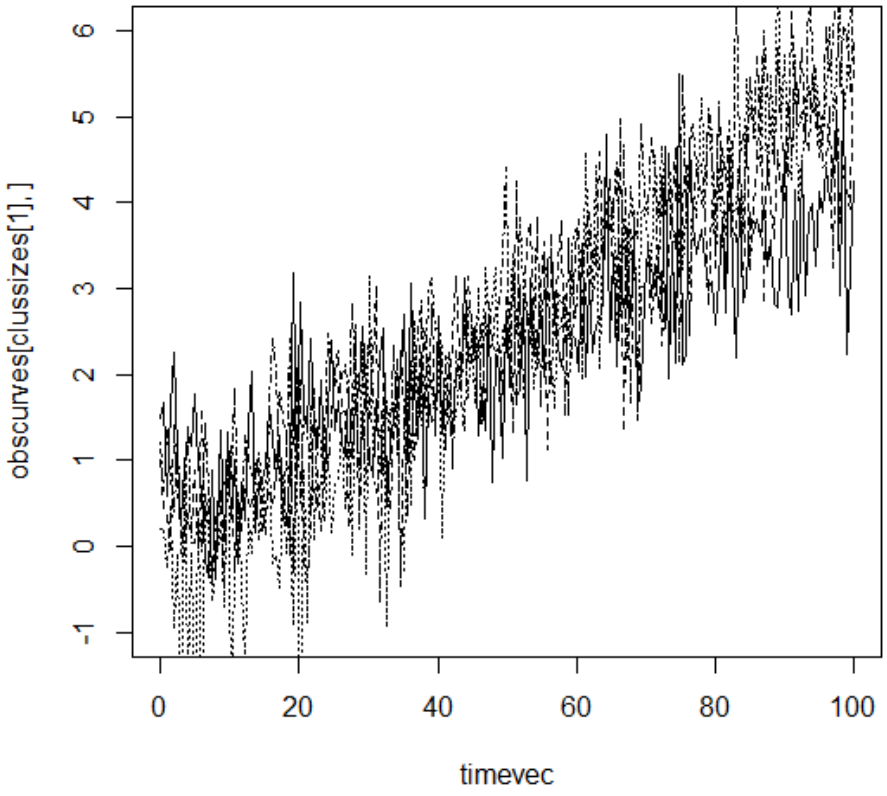

Figure 4.4: Observed curve with Ornstein-Uhlenbeck Process

4.1 Weight functions for functional data

It is important to raise the question whether any of the variables has the tendency to influence our clustering result in any way. In order to avoid favoring any variable type Huang (1998) used weight in his study of K means algorithms for clustering large data sets with categorical variables. The weight was applied only to the categorical variables in this instance. The values of the weight ranges from 0.0 to infinity depending on the nature of the data.[23]. We will consider weighted functions for the functional data in this work, as in the work done by Hendrickson and Hitchcock, there were cases when the functional variable dominated the clustering [32]. We will look at two weight the functions, the first being the inverse variance weight. 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 the sample variance of all $y_i(t) - y_j(t)$ values such that $\sum_i w_i = 1$. The inverse variance weight function puts more weight if curves are more spread apart and less weight on areas where the curves are less spread apart (Chen et al. 2014). The second weight function seeks to minimizes the coefficient of variation of the resulting squared distance between observed functions by means of iterative procedure (CV-optimal

Weight) . This weight function is defined as

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

where $\mathbf{b}_w(t) = [b_{w1}(t), ..., b_{wKw}(t)]^T$ denotes a K_w- dimensional spline basis and $\mathbf{q}$ is the vector of associated spline coefficients [24].

We structured our study in such a way that in some settings we wanted the simulated data to have large distance between clusters. In other we had small distances between clusters. This will indicate to us the performance of the extended Gower and weighted functions under various data composition. In some settings the variables had the same mean for each cluster while in some settings all variables had different means for each cluster. There are four clusters in each data and we changed the cluster sizes for each simulation setting. We considered the following cluster sizes:

- 25 objects in each cluster
- 33 objects in cluster 1,2 and 3 and 1 objects in cluster 4
- We also chose 10 objects in cluster 1, 20 objects in cluster 2, 30 objects in cluster 3 and 40 objects in cluster 4

For each of the 15 simulation settings, we simulated 1000 data sets for each combination of parameter setting and also for each considered cluster sizes. We calculated the rand index, the adjusted rand index as well as the

mean and standard error for each setting. Table 4.1 and 4.2 show the various simulation settings

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	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1.75$
6	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1$
7	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1$
8	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1.75$

We discovered the weights improved the clustering solution using the extended Gower coefficient in most of the settings. The rand index is the largest when using weights in the extended Gower coefficient indicating that the weighted extended Gower coefficient produce the best clustering as compared to the equally weighted L_2 distance.

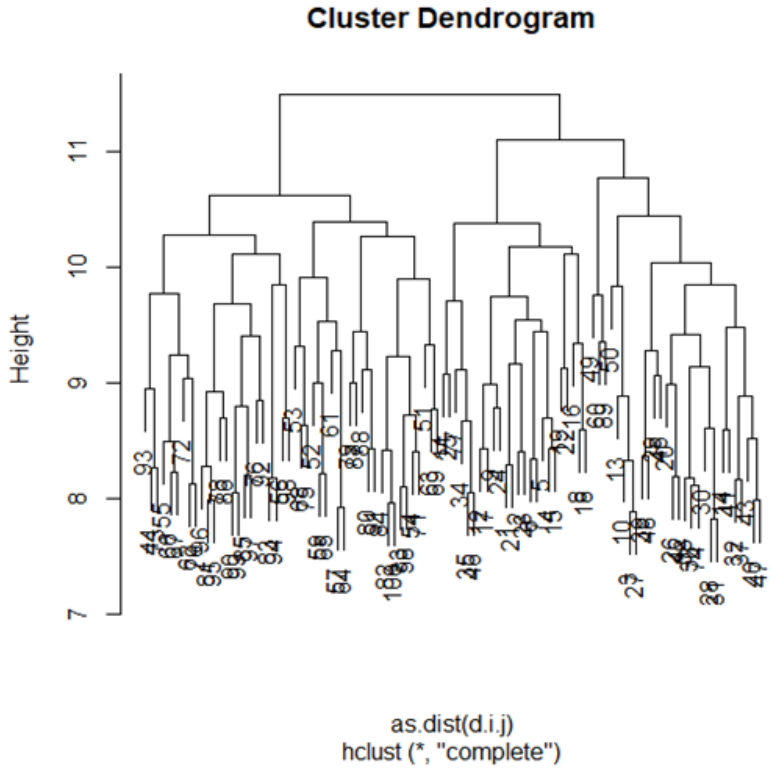

Figure 4.5: A clustering dendrogram showing the extended Gower coefficient.

In some cases the inverse-variance weight performed better while in other

cases the CV-optimal weight proved superior. In other instances the weight functions had no effect on the clustering solution as the average Rand Indices were the same for the equally weighted extended Gower coefficient and the weighted Gower coefficient. It is also observed that the weighted functions produced smaller average rand indices than the equally weighted Gower coefficient.

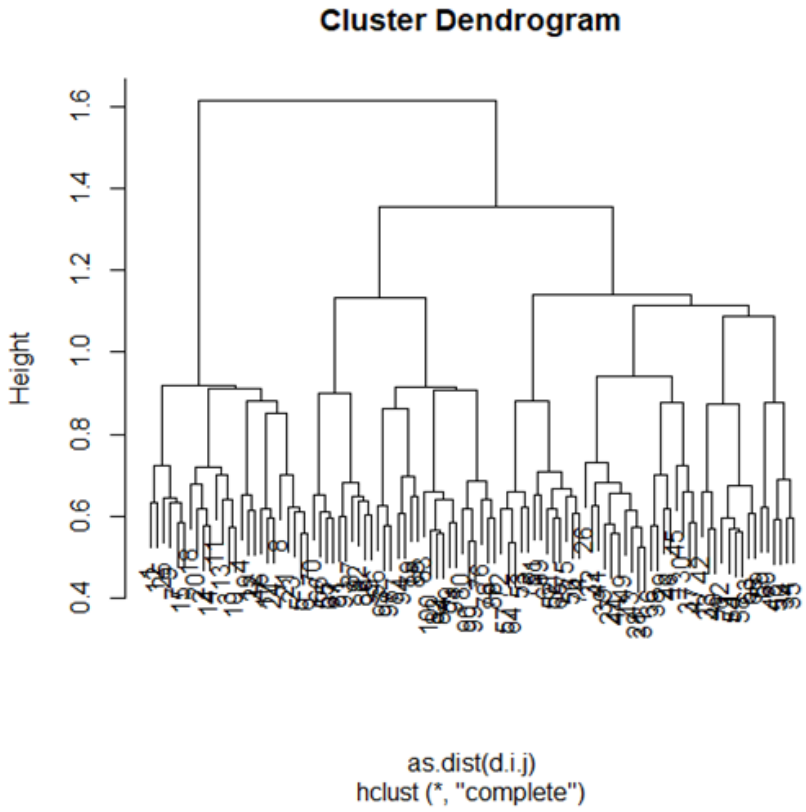

Figure 4.6: A clustering dendrogram showing the extended Gower coefficient with weighted L2.

We make the observation that the Inverse-variance weight proved superior

to the CV-optimal weight when there is a small distance between clusters. That is the value of $\sigma = 1.75$ which indicates small distance between clusters in the Ornstein-Uhlenbeck process. The average rand index in this case was larger for the inverse variance weight.This is true for all but few settings where the average rand index was slightly higher for the CV-optimal weight. In one instance when the cluster sizes are 25 in each cluster the difference was 0.0428. Also when the cluster sizes were 33,33,33 and 1, the CV optimal weight performed better with a difference of 0.0354. When there is a large distance between the clusters of the functional data, that is $\sigma = 1$, the original extended Gower coefficient performed better than the weighted functions, as we had larger values for the average Rand indices. This is the case for most of the settings involving the inverse variance weight. In some settings however, the weight functions had no influence on the extended Gower coefficient. The Rand indices for the extended Gower coefficient and the weighted functions remained unchanged. In general, when there is a large distance between clusters, the extended Gower coefficient without applying weights to the functional variable performed just as well or better than the weighted functional variable. No particular cluster size allocation dominated the outcome of the clustering solution. The average Rand index was higher in some settings with equal cluster sizes, 25 objects in each cluster. In other settings cluster size 33,33,33 and 1 in each cluster produced higher values of

the average Rand index. Likewise, cluster size 40,30,20 and 10 also produced some higher Rand index values in some settings. In some settings where the equally weighted functions was superior in performance to the inverse variance weight (setting 6 ,10 12), we had the categorical variables coming from equiprobable categories though the continuous and the directional variables had different means. In both cases there was large distance between the clusters. In the other settings (settings 2,4,7) where equally weighted L_2 produced higher rand indices, the categorical variables contained one dominant category with probability vectors (0.8,0.05,0.05,0.05,0.05). In these cases the means for the continuous and directional variables were all different though both cases had large separation between clusters ($\sigma = 1$). In only one case do we have the equally weighted distance perform better than the CV-optimal weight(setting 13a). There were 25 objects in each cluster under this setting.

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	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\10000\\15000\\20000\end{bmatrix}$	$\begin{bmatrix}500\\5500\\10500\\15500\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1.75$
10	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
11	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1.75$
12	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1$
13	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1$
14	$\begin{bmatrix}(0.8,0.05,0.05,0.05,0.05)\\(0.05,0.8,0.05,0.05,0.05)\\(0.05,0.05,0.8,0.05,0.05)\\(0.05,0.05,0.05,0.05,0.8)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.1\\0.2\\0.3\end{bmatrix}$	$\sigma = 1.75$
15	$\begin{bmatrix}(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\\(0.2,0.2,0.2,0.2,0.2)\end{bmatrix}$	$\begin{bmatrix}5000\\5500\\6000\\6500\end{bmatrix}$	$\begin{bmatrix}500\\1000\\1500\\2000\end{bmatrix}$	$\begin{bmatrix}0\\0.5\\1\\1.5\end{bmatrix}$	$\sigma = 1.75$

5 DISCUSSION

This work focuses on clustering mixed data involving continuous, categorical, directional and functional variables. We presented a method the creates an extension of the Gower coefficient. We applied hierarchical clustering algorithms to produce the clustering solution. Our focus in this thesis was on weighting the functional data.The application of judiciously chosen weight to the functional data is done to avoid the functional data from influencing the clustering solution. We considered the inverse variance weight and a weight function called the CV-optimal weight which is based on minimizing the coefficient of variation of the squared distance between functional observations. We first used the inverse variance weighting method on a real data set, the buoy data. The Rand index was calculated for equally weighted and the weighted settings.This index gives a proportion of pairs of objects that have been correctly clustered in the same group or correctly clustered into different groups. The Rand Indices were higher for the weighted functional data compared to the equally equally weighted functional data. The clustering solutions based on the weighted data proved superior as the clustering solution classified the buoys into regions relatively close geographically to their "true regions".

In the simulation study, the weighting procedure was shown to improve performance of the extended Gower coefficient under different data settings.

The improvement is quite dramatic compared to the standard unweighted approach. In some settings however, there was no change in performance, as the Rand indices were the same for both the weighted and the unweighted data. This occurred mostly when there were large distances between clusters of the functional variable. The size of the cluster also influenced the average adjusted Rand indices. When cluster sizes were different there were differences in the values of the Average Rand indices though the size of the clusters did not inhibit significantly the work of the influence of the weight functions. In some settings the original unweighted performed better than the weighted data in adjusted Rand indices comparisons. This occurred mostly when the inverse variance weight was applied. One setback with the CV-optimal weight function is that when the size of observed curves is too large it takes a lot of time to compute and may run into high dimensionality problems.

We observed that when there is a small distance between clusters of the functional variables and large separation between the directional variables, the inverse variance weight performed better than the CV-optimal weight. In cases where the equally weighted functional data achieved superiority over the weighted data (inverse variance weight), there were large distances between clusters of the functional data. Also in the setting where no one particular method performed better than the other there was a large distance

clusters of the functional variables and large separation between clusters of the directional variable.

As future work we could look into applying reasonable weight functions on all variables. We could also simulate data containing different types of signal functions. The signal functions used for the simulation study is is a mixture of increasing and decreasing curves. We could look into data containing strictly decreasing functions or a mixture of several functions. In the future, we also expect research on mixed data clustering to continue to be driven by the development of algorithms to handle mixed data.

BIBLIOGRAPHY

[1] Charu C.Aggarwal, Chandan K. Reddy, *Data Clustering Algorithms and Applications*. 13:978-1-4665-58222-9

[2] Adil Fahad, Najlaa Alshatri., Zahir Tari, et. al, *A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis*. Issue No. 03 - Sept. (2014 vol. 2). ISSN: 2168-6750. pp: 267-279.

[3] M. Anderberg. *Cluster Analysis for Applications.* Academic Press, 1973.

[4] Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data.* Prentice Hall

[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 Sons, 2011.

[7] Havard Business review. *hbr-from-data-action-107218*

[8] Cuesta-Albertos, J.A. Gordaliza A., and Matran C. 1997. Trimmed k-means: An attempt to robustify quantizers. *Annals of Statistics*, 25(2), 553-576.

[9] 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.

[10] Jerome H. Friedman and Jacqueline J. Meulman. Clustering objects on subsets of attributes.*J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(4):815–849, 2004.

[11] Kettenring J. R (2006). The practice of clusters Analysis.*Journal of Classification* 23:3-30.

[12] 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.

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

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

[15] Manabu Ichino and Hiroyuki Yaguchi. Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Trans. Systems Man Cybernet.*, 24(4):698–708, 1994

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

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

[18] 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.

[19] Lynette Hunt and Murray Jorgenson. Mixture model clustering using the multimix program. *Australia New Zealand Journal of Statistics*, 41(2):153–171, 1999.

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

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

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

[23] 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.

[24] Huaihou Chen; Philip T Reiss; Thaddeus Tarpey ISSN: 0006-341X; DOI: 10.1111/biom.12161 Biometrics. , 2014, Vol.70(3), p.516-525 Retrieved from: `https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12161`

[25] Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[26] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proceedings of the 26th International Conference on Machine Learning*. Montreal, Canada, June 2009

[27] Piegl L., Tiller W. (1995) B-Spline Basis Functions. In: The NURBS Book. Monographs in Visual Communications. Springer, Berlin, Heidelberg

[28] 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.

[29] FAHAD et al. Survey of Clustering Algorithms for Big Data

[30] Nguyen Hung Son. - Data cleaning and Data Preprocessing.Retrieved from: https://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

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

[32] JeanMarie Hendrickson and David Hitchcock. Methods for Clustering Mixed Data, 2014.

[33] Ordonez, C. (2003). Clustering binary data streams with K-means. *In 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.*

APPENDIX

RAND INDICES ALONG WITH ADJUSTED RAND INDICES

Table .1: Setting 1: Adjusted Rand and Rand Comparisons.

Simulation 1a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 1b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 1c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000

Table .2: Setting 2: Adjusted Rand and Rand Comparisons.

Simulation 2a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.6206	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 2b		
Extended Gower	0.7011	0.8758
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 2c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.9167	0.9655
CV-optimal Weight	1.0000	1.0000

Table .3: Setting 3: Adjusted Rand and Rand Comparisons.

Simulation 3a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.9100	0.9612
CV-optimal Weight	1.0000	1.0000
Simulation 3b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 3c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000

Table .4: Setting 4: Adjusted Rand and Rand Comparisons.

Simulation 4a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.9100	0.9612
CV-optimal Weight	1.0000	1.0000
Simulation 4b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.8989	0.9628
CV-optimal Weight	1.0000	1.0000
Simulation 4c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.8589	0.9412
CV-optimal Weight	1.0000	1.0000

Table .5: Setting 5: Adjusted Rand and Rand Comparisons.

Simulation 5a	Adjusted Rand Index	Rand Index
Extended Gower	0.4500	0.7857
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	0.7369	0.9020
Simulation 5b		
Extended Gower	0.6692	0.8554
Inverse variance Weight	0.9110	0.9620
CV-optimal Weight	0.8677	0.9444
Simulation 5c		
Extended Gower	0.7524	0.8980
Inverse variance Weight	0.9856	0.9941
CV-optimal Weight	0.7074	0.8745

Table .6: Setting 6: Adjusted Rand and Rand Comparisons.

Simulation 6a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.4209	0.7721
CV-optimal Weight	1.0000	1.0000
Simulation 6b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.3684	0.7341
CV-optimal Weight	1.0000	1.0000
Simulation 6c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.4386	0.7154
CV-optimal Weight	1.0000	1.0000

Table .7: Setting 7: Adjusted Rand and Rand Comparisons.

Simulation 7a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.8011	0.9266
CV-optimal Weight	0.9731	0.9901
Simulation 7b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.5426	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 7c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.6579	0.8580
CV-optimal Weight	1.0000	1.0000

Table .8: Setting 8: Adjusted Rand and Rand Comparisons.

Simulation 8a	Adjusted Rand Index	Rand Index
Extended Gower	0.2776	0.7246
Inverse variance Weight	0.9226	0.9715
CV-optimal Weight	0.4345	0.7558
Simulation 8b		
Extended Gower	0.4730	0.7731
Inverse variance Weight	0.9311	0.9707
CV-optimal Weight	0.7541	0.8964
Simulation 8c		
Extended Gower	0.2953	0.6594
Inverse variance Weight	0.9608	0.9836
CV-optimal Weight	0.6965	0.8745

Table .9: Setting 9: Adjusted Rand and Rand Comparisons.

Simulation 9a	Adjusted Rand Index	Rand Index
Extended Gower	0.4012	0.7448
Inverse variance Weight	0.4751	0.7861
CV-optimal Weight	0.5562	0.8289
Simulation 9b		
Extended Gower	0.4147	0.7364
Inverse variance Weight	0.6711	0.8596
CV-optimal Weight	0.6021	0.8343
Simulation 9c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.6579	0.8580
CV-optimal Weight	1.0000	1.0000

Table .10: Setting 10: Adjusted Rand and Rand Comparisons.

Simulation 10a	Adjusted Rand Index	Rand Index
Extended Gower	1.0000	1.0000
Inverse variance Weight	1.0000	1.0000
CV-optimal Weight	1.0000	1.0000
Simulation 10b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.4408	0.7529
CV-optimal Weight	1.0000	1.0000
Simulation 10c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.8784	0.9493
CV-optimal Weight	1.0000	1.0000

Table .11: Setting 11: Adjusted Rand and Rand Comparisons.

Simulation 11a	Adjusted Rand Index	Rand Index
Extended Gower	0.4508	0.7832
Inverse variance Weight	0.8136	0.9307
CV-optimal Weight	0.8203	0.9339
Simulation 11b		
Extended Gower	0.6307	0.8475
Inverse variance Weight	0.8667	0.9428
CV-optimal Weight	0.8252	0.9428
Simulation 11c		
Extended Gower	0.7311	0.8915
Inverse variance Weight	0.8457	0.9366
CV-optimal Weight	0.4721	0.7721

Table .12: Setting 12: Adjusted Rand and Rand Comparisons.

Simulation 12a	Adjusted Rand Index	Rand Index
Extended Gower	0.9731	0.9901
Inverse variance Weight	0.3456	0.7457
CV-optimal Weight	1.0000	1.0000
Simulation 12b		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.5056	0.7869
CV-optimal Weight	0.9432	0.9758
Simulation 12c		
Extended Gower	1.0000	1.0000
Inverse variance Weight	0.2553	0.6941
CV-optimal Weight	1.0000	1.0000

Table .13: Setting 13: Adjusted Rand and Rand Comparisons.

Simulation 13a	Adjusted Rand Index	Rand Index
Extended Gower	0.4035	0.7780
Inverse variance Weight	0.4249	0.7764
CV-optimal Weight	0.3206	0.7279
Simulation 13b		
Extended Gower	0.2795	0.67778
Inverse variance Weight	0.2089	0.6721
CV-optimal Weight	0.4531	0.7743
Simulation 13c		
Extended Gower	0.2284	0.6697
Inverse variance Weight	0.3156	0.7111
CV-optimal Weight	0.2842	0.6770

Table .14: Setting 14: Adjusted Rand and Rand Comparisons.

Simulation 14a	Adjusted Rand Index	Rand Index
Extended Gower	0.4396	0.7885
Inverse variance Weight	0.7233	0.8976
CV-optimal Weight	0.4841	0.7919
Simulation 14b		
Extended Gower	0.3803	0.7285
Inverse variance Weight	0.6339	0.8438
CV-optimal Weight	0.7196	0.8792
Simulation 14c		
Extended Gower	0.0632	0.6026
Inverse variance Weight	0.6112	0.8410
CV-optimal Weight	0.6020	0.8430

Table .15: Setting 15: Adjusted Rand and Rand Comparisons.

Simulation 15a	Adjusted Rand Index	Rand Index
Extended Gower	0.1573	0.6511
Inverse variance Weight	0.5797	0.8325
CV-optimal Weight	0.3588	0.7396
Simulation 15b		
Extended Gower	0.4433	0.7642
Inverse variance Weight	0.4586	0.7747
CV-optimal Weight	0.4816	0.7630
Simulation 15c		
Extended Gower	0.3775	0.7372
Inverse variance Weight	0.6375	0.8519
CV-optimal Weight	0.3568	0.7420

VITA

AUGUSTINE OPPONG

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 2012
Professional Experience:	Graduate Teaching Assistant, East Tennessee State University Johnson City, Tennessee, 2016–2018 Health Insurance Claims officer, Kintampo Municipal Hospital Kintampo, Ghana, 2013–2016
Professional Development:	Statistical and Mathematical, Software: SAS, R, SPSS, Minitab Microsoft Office Suite: MS Access VBA, Word, Excel, PowerPoint, Outlook