Dissimilarity Measures for Histogram-valued Data and Divisive Clustering of Symbolic Objects

by

Jaejik Kim

(Under the direction of Lynne Billard)

Abstract

Contemporary datasets are becoming increasingly larger and more complex, while techniques to analyse them are becoming more and more inadequate. Thus, new methods are needed to handle these new types of data. This study introduces methods to cluster histogram-valued data. However, histogram-valued data are difficult to handle computationally because observations typically have a different number and length of subintervals. Thus, a transformation for histogram data is proposed as a technique for handling them more easily computationally. From this technique, three new dissimilarity measures for histogram data are proposed. Then, how the monothetic clustering algorithm based on Chavent (1998, 2000) can be extended to histogram data is shown, and a polythetic clustering algorithm for symbolic objects is developed (based on all p variables). Validity criteria to aid in the selection of the optimal number of clusters are described and verified by some simulation studies. The new methodology is illustrated on a large dataset collected from the US Forestry Service.

Index words: Symbolic data, Histogram-valued data, Dissimilarity measure, Monothetic algorithm, Polythetic algorithm, Validity

Dissimilarity Measures for Histogram-valued Data and Divisive Clustering of Symbolic Objects

by

Jaejik Kim

B.A., SungKyunKwan University, 1999

M.S., SungKyunKwan University, 2001

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2009

Dissimilarity Measures for Histogram-valued Data and Divisive Clustering of Symbolic Objects

by

Jaejik Kim

Approved:

Major Professor:	Lynne Billard
Committee:	T. N. Sriram William P. McCormick Jaxk Reeves Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2009

Dedication

To my parents, my wife, and my daughter

Acknowledgments

I would first like to thank my major advisor Dr. Lynne Billard. Dr. Lynne Billard for all of her guidance, and patience, in the writing and research of this dissertation, has made my graduate research exciting and interesting.

I would also like to thank my committee members, Dr. T.N. Sriram, Dr. William P. McCormick, Dr. Jaxk Reeves, and Dr. Lynne Seymour for their support, careful reading and suggestions. I also would like to show my appreciation to the faculty and staff of the department. They provided the best place ever and let me have a chance to pursue a higher degree. Many thanks to Dr. Jin-Hong Park and Jaejun Lee for their advice and for being such good friends.

Finally and most importantly, I thank my wife and my family for their support, their sacrifices, and their unconditional love.

Table of Contents

Page

Chapter

Appendix

List of Figures

List of Tables

Chapter 1

Introduction

Nowadays with contemporary computer capacity, the size of datasets is rapidly increasing in many areas with an increase of both the number of objects and variables. Under this trend, methods of summarizing and extracting information from large datasets are becoming more and more important.

One approach to handling very large datasets is to aggregate the data in some meaningful way. When we are interested in classes or groups of individuals, the data for individuals can be aggregated as classes or groups. The form of the aggregated data can be a range, list, and distribution, etc. These types of data are called symbolic data; Diday (1987), Bock and Diday (2000), and Billard and Diday (2006). In order to analyse symbolic data, we need new methods which are different from classical methods in many aspects, and in particular in cluster analysis as well. Our focus is on clustering methods for symbolic data, particularly histogram data.

For clustering histogram data, we need new dissimilarity measures and clustering algorithms. Gowda and Diday (1991a, 1992) suggested similarity and dissimilarity measures for multi-valued and interval-valued variables and a hierarchical agglomerative clustering algorithm. Also, Ichino and Yaguchi (1994) proposed dissimilarity measures for multi-valued and interval-valued data and extensions to Minkowski distances.

An agglomerative clustering method starts with clusters which are equal in number to the number of objects. In other words, each cluster has one object at the first stage. It merges successively two clusters until all objects are in a cluster. In contrast, a divisive method starts with a cluster containing all objects, and then bisects successively into two clusters until there is one object in each cluster. That is, a divisive clustering is the reverse of an agglomerative clustering. Although the divisive method is less popular than the agglomerative method, Kaufman and Rousseeuw (1990) indicated that it has the advantage which it shows the main structure in the data. Chavent (1998, 2000) suggested a hierarchical monothetic divisive clustering algorithm for interval data. A monothetic algorithm uses a single variable to bisect a cluster at each stage. In contrast, a polythetic algorithm uses all the variables at each stage. MacNaughton-Smith et al. (1964) proposed a polythetic method using an iterative procedure based on an average dissimilarity between an object and a group of objects.

While an interval-valued random variable is defined by the lower and upper limits $[a, b)$, a histogram-valued random variable is defined by a finite number of non-overlapping subintervals and relative frequencies $\{[a_k, a_{k+1});\ p_k,\ k = 1, \ldots, v\}$. A histogram-valued variable is more informative than an interval-valued variable because the former provides estimates of the shape and location for the distribution but the latter gives only lower and upper limits. Therefore, clustering for histogram-valued data gives more precise outcomes than that for interval-valued objects. However, such methods are not available since there are no existing dissimilarity measures for histogram-valued data. Moreover, it is not easy to handle histogram data in a computer because each observation has different lengths of subintervals as well as a different number of subintervals.

In general, we do not have any prior information about the number of clusters in the data. Moreover, since the clusters are often indistinguishable in the aspect of dissimilarity measures and different dissimilarity measures often lead to different clustering outcomes, it is not easy to detect the optimal number of clusters. To solve this, many different cluster validity indexes have been proposed, such as Dunn's index (1974), Davis-Bouldin's index (1979), and Xie-Beni's index (1991), among others. However, since these indexes were developed for classical or fuzzy datasets, they would not work well for symbolic data.

In this study, we develop clustering methodology for histogram-valued data. In Chapter 2, dissimilarity measures for interval data and the monothetic algorithm are reviewed. In

Chapter 3, we propose a transformation of histogram data to enable the new methodologies to be handled more easily computationally. From this technique, we develop dissimilarity measures for histogram data based on the Gowda-Diday and Ichino-Yaguchi measures and cumulative relative frequencies. In addition, we show how the monothetic algorithm based on Chavent(1998, 2000) can be extended to histogram data and propose a polythetic algorithm to produce new divisive hierarchical clusterings for symbolic objects. Cluster validity indexes that are concerned with determining the optimal number of clusters are discussed in Chapter 4; simulation studies are also executed to study their properties. The methodology proposed herein is illustrated on a forestry cover type dataset of 581,012 observations, in Chapter 5.

Chapter 2

Literature Review

In this chapter, we review the literature on dissimilarity/distance measures and clustering methods for interval-valued data. In Section 2.1, methods to deal with large datasets such as data squashing, boosting, and data mining, etc., are reviewed, and the difference between these methods and symbolic data is explained. The definitions of symbolic data and their various examples are presented in Section 2.2. The clustering methodologies for interval-valued data are reviewed in Section 2.3 and 2.4. In Section 2.3, various similarity and dissimilarity measures for interval-valued data are introduced. In Section 2.4, the monothetic algorithm for interval-valued data proposed by Chavent (1988, 2000) is explained and this algorithm is illustrated using the Ruspini (1970) data.

2.1 Large DataSets

Datasets are becoming larger and larger and more complex with contemporary computer capacity. On the contrary, techniques to analyse them have been overwhelmed by the pace of the data collection. It is evident that classical methods are often unable to handle very large datasets. Thus, new methods to analyse them such as data squashing, boosting, and data mining, etc., have been introduced in many studies.

Data squashing, introduced by DuMouchel et al. (1999), produces a small sample from a large dataset by aggregating the data into clusters and taking from each cluster a sample with similar characteristics as determined by pseudodata points and weights. This sample represents the large original dataset. Madigan et al. (2002) extends the moment matching of DuMouchel et al. to likelihood-based squashing. In addition, Owen (2003) proposes data squashing by empirical likelihood, but shows that data squashing does not always give good results, and suggests the methods need to be applied to more extensive datasets before conclusions can be drawn. Owen also considers boosted classification and decision trees along the lines of multiple adaptive regression trees for constructing boosted tree classifications. Inatani and Suzuki (2002) combine data squashing and boosting methods in the computation of outlier detection. Kaufman and Rousseeuw (1990) use the CLARA (Clustering LARge Applications) algorithm to reduce the dataset to a sample. Another technique that deals with large datasets is data mining; this method is interested in discovering patterns and extracting knowledge from the dataset.

Symbolic data methods produce smaller datasets by the summarization of very large datasets. It seems that the symbolic data methodology is similar to data squashing from the point of view of producing smaller datasets. However, the symbolic data method as a result of aggregation is different from data squashing because the former uses all the original data but the latter uses just a sample. Since data squashing uses a sample, it can use standard statistical methods to analyse a resulting data set. In contrast, symbolic data require new statistical techniques because they are new types of data such as lists, intervals, histograms, models, etc. In addition, some problems in data mining and knowledge discovery lead naturally to symbolic data.

The different methods being proposed in this section will have a key role to play in differing contexts, and equally each overlaps the others with common issues serving all fields. Symbolic data methods also will help to play a role in analyzing large datasets. The concept of symbolic data was first introduced by Diday (1987), and are described in Bock and Diday (2000), and Billard and Diday (2006).

2.2 Symbolic Data

In the past decade, information technology has developed remarkably. Consequently, large and complex datasets are now common due to routine collection of systematically generated data. For example, databases of credit card companies record all transactions for all card users, and mobile companies keep the records of all cell phone calls. A huge amount of records are routinely generated everyday. These datasets have billions of observations and hundreds of variables. Since datasets are too large to be stored in the primary memory of a computer, it is difficult to calculate statistics from large datasets by the usual statistical algorithms and methodologies.

One approach is to summarize large datasets in a meaningful way. In the mobile company example, a summary of the calls per user can be made instead of hundreds as specific calls for each user over time. One such summary format could be a range of time for each call (e.g., 2 – 67 minutes); or, it could be a list of received calls (e.g., Home, Mary, Mother,...); or, it could be by type and calling time (e.g., {received call, 2 – 35 minutes}, {dialed call, 5 – 67 minutes}); or, it could be a histogram by calling time and relative frequencies for the number of calls (e.g., {0 – 10 minutes, 0.5}, {10 – 20 minutes, 0.3},{20 – 30 minutes, 0.2}); or, etc. Like these examples, the summarized data can be ranges, lists, histograms, distributions, and models, etc., and these types of data are called symbolic data. These symbolic datasets have a manageable size in a computer.

Notation 2.1 *For the random variable $Y_j,\ j = 1, \ldots, p$, a symbolic value or realization $i = 1, \ldots, n$, will be denoted by y_{ij}. Also, a symbolic object (or observation) will be denoted by $\mathbf{y}_i = (y_{ij},\ j = 1, \ldots, p) \in \Omega$.*

Based on the attendant research questions, classical large datasets are aggregated into classes or groups. The outcomes aggregated by classes or groups are symbolic objects $\mathbf{y}_i,\ i = 1, \ldots, n \in \Omega$, where $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Each symbolic object consists of symbolic values $y_{ij},\ j = 1, \ldots, p$, and symbolic values can be lists, intervals, histograms, and models, etc. For example, suppose we have a dataset including individual records for pitchers such as innings pitched and earned runs. Then, we might be interested in the pitching performance of baseball teams rather than individual players. To solve this question, first of all, we would aggregate the data over individual players who make up each team. The summarized data

Table 2.1: Bird colors.

Bird	Major Colors
Anhinga	{black, green}
Bananaquit	{black, white, yellow}
Blue Jay	{blue, gray, black}
Common Redpoll	{red, gray, black, brown}
Budgerigar	{green, blue, white, yellow }

would be intervals and histograms. For example, suppose that the data are aggregated as intervals on Y_1 =innings pitched and Y_2 =earned runs, and there are 30 baseball teams. Then, symbolic objects $\mathbf{y}_i$ are baseball teams and can be denoted by $\mathbf{y}_i \in \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_{30}\}$, where $\mathbf{y}_i = \{y_{i1} = [a_{i1}, b_{i1}),\ y_{i2} = [a_{i2}, b_{i2})\}$. The symbolic objects can be referred to as categories or classes. In this example, categories are baseball teams. Therefore, the data for individuals can be reduced to the data for teams by the research question.

Although symbolic data analysis is a method to deal with large datasets, symbolic data can exist regardless of dataset size. For example, suppose we are interested in bird colors. Most birds have two or more colors. Thus, since a classical value takes a single value, it cannot express colors of a bird. In contrast, since symbolic data have internal structure, it is possible to express the colors as shown in Table 2.1. Thus, e.g., a blue jay has the three colors blue, gray, and black. Table 2.1 does not come from a large dataset. In this case, the type of symbolic data is a list of single values, and it is called multi-valued data.

Definition 2.1 *A **multi-valued** random variable Y takes one or more values from the list of values in its domain $\mathcal{Y}$.*

Definition 2.2 *An **interval-valued** random variable Y takes values in an interval $[a, b] \subset \Re$, where $a \leq b$, $a, b \in \Re$. The interval can be closed or open.*

Often, symbolic data naturally arise. For example, a pulse rate usually fluctuates over the interval. (e.g., $[64 \pm 2]$); or, in a weather forecast, daily temperature can be usually expressed

Table 2.2: Histograms for weight by age-groups.

Age	Weight
20s	{[70,96),0.08; [96,108),0.24; [108,120),0.30; [120,144),0.30; [144,160),0.08}
30s	{[100,116),0.08; [116,124),0.40; [124,132),0.24; [132,140),0.24; [140,150),0.04}
40s	{[110,135),0.18; [135,145),0.20; [145,155),0.42; [155,165),0.14; [165,185),0.06}
50s	{[100,126),0.10; [126,138),0.20; [138,150),0.26; [150,162),0.28; [162,190),0.16}
60s	{[125,144),0.18; [144,160),0.60; [160,168),0.16; [168,180),0.06}

by the lowest and highest temperature in a day (e.g., [45,78]), etc. In such examples, classical methods usually use midpoints for each interval. However, it causes a loss of information especially since the internal variation is ignored. Consider the following simple example where we have three samples on the random variable Y =weight. Suppose $Y_1 = 130$, $Y_2 = [127, 133]$, and $Y_3 = [124, 136]$. Also, assume that an interval has an uniform distribution. Then, means for all three samples are the same, i.e., $\bar{Y}_1 = \bar{Y}_2 = \bar{Y}_3 = 130$, but they have different internal variations, i.e., $Var(Y_1) = 0$, $Var(Y_2) = 3$, $Var(Y_3) = 12$. Classical analysis using midpoint values give the same results for the three samples. In contrast, symbolic analysis can provide more informative results because it considers the internal variation for each sample.

Definition 2.3 *Let Y be a random variable that takes values on a finite number of non-overlapping intervals $\{[a_k, a_{k+1}),\ k = 1, \ldots, v\}$ with relative frequencies p_k corresponding to each subinterval, where $a_k \leq a_{k+1}$. Then, Y is called a* ***histogram-valued*** *random variable. The i^{th} observation y_i for a histogram-valued random variable is given by*

$$y_i = \{[a_{ik}, a_{i,k+1}),\ p_{ik};\ k = 1, \ldots, v_i\},$$

where $\sum_{k=1}^{v_i} p_{ik} = 1$.

A type of data that is more informative than interval-valued data is histograms. Histogram-valued data consist of a finite number of non-overlapping subintervals and

Table 2.3: Distribution of monthly water usage for households.

Household	Water usage (gallon)
A	Normal ($\mu = 2730,\ \sigma = 347$)
B	Normal ($\mu = 1882,\ \sigma = 672$)
C	Normal ($\mu = 3472,\ \sigma = 245$)
D	Normal ($\mu = 4324,\ \sigma = 781$)
E	Normal ($\mu = 2320,\ \sigma = 145$)

relative frequencies corresponding to each subinterval as shown in Table 2.2. Since histograms with well defined subintervals can be a good density estimate, it is a useful way to summarize large datasets.

Furthermore, in the symbolic data context, histogram-valued observations could be specified distributions. Distribution data could be known parametric distributions such as normal, exponential, and gamma distributions; or they could be empirical distributions with the parameter values estimated from data. Table 2.3 shows an example with distributions as symbolic data. Suppose monthly water usage for households has a normal distribution. Then, parameters of normal distributions are estimated from the original data, and water usage for each household can be summarized by the distributions.

Models also can be used as symbolic data. For example, suppose a trend for the price of a stock of each company has autoregressive moving average (ARMA) models as shown in Table 2.4. Then, in this case the time series data for the price of a stock of each company can be summarized by ARMA models with parameters, and we can use these models as symbolic data for each company.

In summary, the form of symbolic data can be lists, intervals, distributions, and models, etc. While classical data on p random variables are expressed by single points in p-dimensional space, symbolic data on p random variables are p-dimensional hypercubes, or a Cartesian product of p distributions, broadly defined. In addition, each symbolic observation has

Table 2.4: Model of the price of a stock for companies.

Company	Model
Company 1	AR(2); $\phi_1 = 0.7,\ \phi_2 = -0.2$
Company 2	ARMA(1,1); $\phi_1 = 0.3,\ \theta_1 = 0.1$
Company 3	AR(1); $\phi_1 = 0.5$
Company 4	AR(2,1); $\phi_1 = 0.3,\ \phi_2 = -0.1,\ \theta_1 = 0.2$

internal variation, but a classical observation does not. Thus, classical analyses deal with variation between observations only. In contrast, analyses of symbolic data have to explain both the internal variations and the variations between observations. More detailed descriptions of symbolic data along with numerous examples can be found in Billard and Diday (2003, 2006).

2.3 An Overview of Similarity and Dissimilarity Measures for Interval-valued Data

Clustering is an exploratory procedure to understand data with complex structure and multivariate relationships. In general, there are two main parts that have important roles in clustering. One is similarity or dissimilarity (or distance) measures, and another is clustering algorithms. Usually, clustering is performed by various criteria such as the smallest distance (single linkage), the farthest distance (complete linkage), Ward criterion (1963), and the minimum within-cluster variance, etc. While a few methods exist which use the observations directly (e.g., Brito 1995), most criteria are based on similarity or dissimilarity measures. Thus, it is very important to define similarity or dissimilarity measures between two objects. In this section, we review similarity and dissimilarity measures for interval-valued data. While classical point values are special cases of symbolic data, these measures are very different for symbolic data than are those for classical data. An extensive summary of classical measures can be found in Gordon (1999), such as the Minkowski metrics, the Canberra metric (Lance and Williams, 1966), the angular separation, and the correlation coefficient, etc.

Before reviewing similarity and dissimilarity measures for interval-valued data, we give definitions of similarity, dissimilarity, and distance measures.

Definition 2.4 *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be any two objects in* Ω*. Then, a* ***similarity measure*** $S(\mathbf{x}, \mathbf{y})$ *between the objects* $\mathbf{x}$ *and* $\mathbf{y}$ *has the following properties:*

$< i >$ $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x})$*;*

$< ii >$ $S(\mathbf{x}, \mathbf{x}) = S(\mathbf{y}, \mathbf{y}) > S(\mathbf{x}, \mathbf{y})$ *for all* $\mathbf{x} \neq \mathbf{y}$*;*

Property $< i >$ shows the symmetric property of similarity measures, and $< ii >$ means that the similarity between the same objects has the largest value.

Definition 2.5 *A* ***dissimilarity measure*** $D(\mathbf{x}, \mathbf{y})$ *between the objects* $\mathbf{x}$ *and* $\mathbf{y}$ *is a measure that satisfies*

(i) $D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x})$*;*

(ii) $D(\mathbf{x}, \mathbf{x}) = D(\mathbf{y}, \mathbf{y}) < D(\mathbf{x}, \mathbf{y})$ *for all* $\mathbf{x} \neq \mathbf{y}$*;*

(iii) $D(\mathbf{x}, \mathbf{x}) = 0$ *for all* $\mathbf{x} \in \Omega$*.*

Property (i) represents the symmetric property of dissimilarity measures. From (ii) and (iii), we can know that the dissimilarity value between the same two objects is zero, and a dissimilarity value is always positive. Typically, the relationship between similarity and dissimilarity measures is inverse functional. For example, similarity measures can be easily transformed to dissimilarity measures by $D(\mathbf{x}, \mathbf{y}) = 1 - S(\mathbf{x}, \mathbf{y})$.

Definition 2.6 *A* ***distance measure*** *has the properties of a dissimilarity measure as defined in (i), (ii), and (iii), and further satisfies*

(iv) $D(\mathbf{x}, \mathbf{y}) = 0$ *implies* $\mathbf{x} = \mathbf{y}$*;*

(v) $D(\mathbf{x},\mathbf{y}) \leq D(\mathbf{x},\mathbf{z}) + D(\mathbf{z},\mathbf{y})$ *for all* $\mathbf{x},\ \mathbf{y},\ \mathbf{z} \in \Omega$.

Property (v) in the definition of a distance measure is called the triangular inequality. Since dissimilarity measures do not satisfy the triangular inequality, unlike distances, dissimilarity cannot be geometrically explained.

As mentioned in Section 2.2, classical data are single points in $p-$dimensional space. Thus, since classical data only have information for location and satisfy the triangular inequality, they can be geometrically explained in $p-$dimensional space. Therefore, the dissimilarity of classical data can be measured by distance measures. In contrast, since symbolic data have information for internal variation as well as location, dissimilarity should explain both location and internal variation of symbolic objects. Thus, Nieddu and Rizzi (2003) indicated that dissimilarities for symbolic data do not always satisfy the triangular inequality.

Finally, for the collection of objects $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in Ω, the $n \times n$ matrix $\mathbf{D}$ with elements $D(\mathbf{y}_i, \mathbf{y}_j),\ i, j = 1, \ldots, n$ is called the *dissimilarity matrix* or *distance matrix.*

There exist many similarity and dissimilarity measures for interval-valued data. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$ be a vector of p interval-valued variables. Then, the interval-valued observation $\mathbf{y}_i$ for i^{th} object is given by

$$\mathbf{y}_i = (y_{ij},\ j = 1, \ldots, p) = \{[a_{ij}, b_{ij}),\ j = 1, \ldots, p\},\ i = 1, \ldots, n. \tag{2.1}$$

Now, we review these similarity and dissimilarity measures and their properties. Gowda and Diday (1991b) proposed a similarity measure for interval-valued data. The *Gowda-Diday similarity measure* between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ is given by

$$S_{GD}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^{p} \big[S_1(y_{i_1j}, y_{i_2j}) + S_2(y_{i_1j}, y_{i_2j}) + S_3(y_{i_1j}, y_{i_2j})\big], \tag{2.2}$$

where for the j^{th} variable, each component of the measure is

$$S_1(y_{i_1j}, y_{i_2j}) = \frac{|b_{i_1j} - a_{i_1j}| + |b_{i_2j} - a_{i_2j}|}{2|\max(b_{i_1j}, b_{i_2j}) - \min(a_{i_1j}, a_{i_2j})|},$$

$$S_2(y_{i_1j}, y_{i_2j}) = \frac{\Lambda_{i_1i_2j}}{|\max(b_{i_1j}, b_{i_2j}) - \min(a_{i_1j}, a_{i_2j})|},$$

where

$$\Lambda_{i_1 i_2 j} = \begin{cases} |\max(a_{i_1 j}, a_{i_2 j}) - \min(b_{i_1 j}, b_{i_2 j})|, & \text{if } \max(a_{i_1 j}, a_{i_2 j}) < \min(b_{i_1 j}, b_{i_2 j}), \\ 0, & \text{otherwise,} \end{cases}$$

i.e., $\Lambda_{i_1 i_2 j}$ is the length of the overlapped interval between $y_{i_1 j}$ and $y_{i_2 j}$; and

$$S_3(y_{i_1 j}, y_{i_2 j}) = 1 - \frac{|a_{i_1 j} - a_{i_2 j}|}{\max_i(b_{ij}) - \min_i(a_{ij})}.$$

The denominator in $S_3(y_{i_1 j}, y_{i_2 j})$ equals the total length spanned by all of the observations of variable Y_j.

As shown in Equation (2.2), the Gowda-Diday similarity measure consists of three components. The first component, $S_1(y_{i_1 j}, y_{i_2 j})$, measures the relative sizes of two objects without referring to common parts between them. The second component, $S_2(y_{i_1 j}, y_{i_2 j})$, indicates the common parts of two objects. The last component, $S_3(y_{i_1 j}, y_{i_2 j})$, measures the relative position of two objects. All three components have normalized values between 0 and 1.

The Gowda-Diday dissimilarity measure for interval-valued data was introduced in Gowda and Diday (1991a). Similarly to the Gowda-Diday similarity measure, the *Gowda-Diday dissimilarity measure* between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ can be defined by

$$D_{GD}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^{p} \big[D_1(y_{i_1 j}, y_{i_2 j}) + D_2(y_{i_1 j}, y_{i_2 j}) + D_3(y_{i_1 j}, y_{i_2 j})\big], \tag{2.3}$$

where for the j^{th} variable, each component of the measure is

$$D_1(y_{i_1 j}, y_{i_2 j}) = \frac{\big||b_{i_1 j} - a_{i_1 j}| - |b_{i_2 j} - a_{i_2 j}|\big|}{|\max(b_{i_1 j}, b_{i_2 j}) - \min(a_{i_1 j}, a_{i_2 j})|},$$

$$D_2(y_{i_1 j}, y_{i_2 j}) = \frac{|b_{i_1 j} - a_{i_1 j}| + |b_{i_2 j} - a_{i_2 j}| - 2\Lambda_{i_1 i_2 j}}{|\max(b_{i_1 j}, b_{i_2 j}) - \min(a_{i_1 j}, a_{i_2 j})|},$$

where

$$\Lambda_{i_1 i_2 j} = \begin{cases} |\max(a_{i_1 j}, a_{i_2 j}) - \min(b_{i_1 j}, b_{i_2 j})|, & \text{if } \max(a_{i_1 j}, a_{i_2 j}) < \min(b_{i_1 j}, b_{i_2 j}), \\ 0, & \text{otherwise,} \end{cases}$$

i.e., $\Lambda_{i_1 i_2 j}$ is the length of the overlapped interval between $y_{i_1 j}$ and $y_{i_2 j}$; and

$$D_3(y_{i_1 j}, y_{i_2 j}) = \frac{|a_{i_1 j} - a_{i_2 j}|}{\max_i(b_{ij}) - \min_i(a_{ij})}.$$

Similarly to Equation (2.2), the Gowda-Diday dissimilarity measure is constituted by three components. The first component, $D_1(y_{i_1j}, y_{i_2j})$, corresponds to the relative size of y_{i_1j} and y_{i_2j} including common parts between them. The $D_2(y_{i_1j}, y_{i_2j})$ indicates the relative content excluding common parts between them. Finally, $D_3(y_{i_1j}, y_{i_2j})$ is a measure of their relative positions.

Gowda and Ravi (1995) introduced a modified version for both Gowda-Diday similarity and dissimilarity measures. They indicated that the similarity and dissimilarity measures for interval-valued data introduced by Gowda-Diday (1991a,b) have several disadvantages as follows: Firstly, when there is no overlapping part between two interval-valued objects, the dissimilarity measure is greater than the similarity measure. Secondly, when the lengths of two intervals are the same, the similarity measure is greater than the dissimilarity measure. Thirdly, the third component in the similarity measure, $S_3(y_{i_1j}, y_{i_2j})$, is just another aspect of that component in the dissimilarity measure, $D_3(y_{i_1j}, y_{i_2j})$. To overcome these disadvantages, Gowda and Ravi (1995) modified both the similarity and dissimilarity measures for interval-valued objects. The *Gowda-Ravi similarity measure* between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ is defined using a sine function as follows:

$$S_{GR}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^{p} \left[S_1^*(y_{i_1j}, y_{i_2j}) + S_2^*(y_{i_1j}, y_{i_2j}) \right], \tag{2.4}$$

where

$$S_1^*(y_{i_1j}, y_{i_2j}) = \sin\left[90 \left(\frac{|b_{i_1j} - a_{i_1j}| + |b_{i_2j} - a_{i_2j}|}{2|\max(b_{i_1j}, b_{i_2j}) - \min(a_{i_1j}, a_{i_2j})|} \right) \right],$$

and

$$S_2^*(y_{i_1j}, y_{i_2j}) = \sin\left[90 \left(1 - \frac{|a_{i_1j} - a_{i_2j}|}{\max_i(b_{ij}) - \min_i(a_{ij})} \right) \right].$$

Unlike the Gowda-Diday similarity measure of Equation (2.2), this measure consists of two components, relative size and position between two interval-valued objects. The first component $S_1^*(y_{i_1j}, y_{i_2j})$ measures the relative size between two objects, and the second component $S_2^*(y_{i_1j}, y_{i_2j})$ indicates the relative position. That is, this modified measure does not consider the component of relative content.

The *Gowda-Ravi dissimilarity measure* between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ is given using a cosine function as follows:

$$D_{GR}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \sum_{j=1}^{p} \left[D_1^*(y_{i_1j}, y_{i_2j}) + D_2^*(y_{i_1j}, y_{i_2j}) \right], \tag{2.5}$$

where

$$D_1^*(y_{i_1j}, y_{i_2j}) = \cos\left[90 \left(\frac{|b_{i_1j} - a_{i_1j}| + |b_{i_2j} - a_{i_2j}|}{2|\max(b_{i_1j}, b_{i_2j}) - \min(a_{i_1j}, a_{i_2j})|} \right) \right],$$

and

$$D_2^*(y_{i_1j}, y_{i_2j}) = \cos\left[90 \left(1 - \frac{|a_{i_1j} - a_{i_2j}|}{\max_i(b_{ij}) - \min_i(a_{ij})} \right) \right].$$

The Gowda-Ravi dissimilarity measure is also constituted by two components, but it uses a cosine function instead of a sine function. By using sine and cosine functions, Gowda-Ravi similarity and dissimilarity measures can overcome the disadvantages of the Gowda-Diday measures mentioned above.

The *Ichino-Yaguchi dissimilarity measure*, proposed by Ichino and Yaguchi (1994), is defined using the Cartesian operators 'join' and 'meet' between two sets. For the interval-valued variable Y_j, the Ichino-Yaguchi dissimilarity measure between objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ is, for variable $Y_j,\ j = 1, \ldots, p$,

$$\phi(y_{i_1j}, y_{i_2j}) = |y_{i_1j} \oplus y_{i_2j}| - |y_{i_1j} \otimes y_{i_2j}| + \gamma(2|y_{i_1j} \otimes y_{i_2j}| - |y_{i_1j}| - |y_{i_2j}|), \tag{2.6}$$

where the Cartesian join $y_{i_1j} \oplus y_{i_2j}$ for interval-valued objects is

$$y_{i_1j} \oplus y_{i_2j} = \big[\min(a_{i_1j}, a_{i_2j}), \max(b_{i_1j}, b_{i_2j}) \big), \tag{2.7}$$

and the Cartesian meet $y_{i_1j} \otimes y_{i_2j}$ for interval-valued observations is

$$y_{i_1j} \otimes y_{i_2j} = \begin{cases} \big[\max(a_{i_1j}, a_{i_2j}), \min(b_{i_1j}, b_{i_2j}) \big), & \text{if } \max(a_{i_1j}, a_{i_2j}) < \min(b_{i_1j}, b_{i_2j}), \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

where $|\cdot|$ is the length of the interval (e.g., $|y_{ij}| = b_{ij} - a_{ij}$), and γ is a prespecified constant between 0 and 0.5 ($0 \le \gamma \le 0.5$). Unlike the Gowda-Diday dissimilarity measure, since the Ichino-Yaguchi measure is not a normalized measure, it has different scales for each variable.

If an unnormalized measure is used, its value might depend on variables with large scales. In order to solve this problem, the total length spanned by observations for variable Y_j is used as a normalized factor. Therefore, the normalized Ichino-Yaguchi measure for Y_j is given by

$$\phi^*(y_{i_1j}, y_{i_2j}) = \frac{\phi(y_{i_1j}, y_{i_2j})}{\max_i(b_{ij}) - \min_i(a_{ij})}. \tag{2.9}$$

This normalized Ichino-Yaguchi measure has a value between 0 and 1.

Malerba et al. (2001) in their empirical study proposed that when $|y_{i_1j} \otimes y_{i_2j}| = 0$, γ be set to 0.5 to prevent nullifying the contribution of the Cartesian meet operator. In addition, they recommended generally to use an intermediate value between 0 and 0.5.

Ichino and Yaguchi (1994) also suggested extensions to Minkowski distances for their dissimilarity measure. The *generalized Minkowski distance* between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ is defined by

$$D_M^q(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \Big[\sum_{j=1}^{p} \phi(y_{i_1j}, y_{i_2j})^q\Big]^{1/q}, \tag{2.10}$$

where $\phi(y_{i_1j}, y_{i_2j})$ is given in Equation (2.6), and $q \geq 1$ is a prespecified order. According to various values of order q, this provides various measures. If order $q = 1$, this becomes the *city block distance* (also called the Manhattan distance); and a generalized Minkowski distance with order $q = 2$ is called the *Euclidean distance*.

If Equation (2.9) is applied to the generalized Minkowski distance, its normalized version can be obtained as follows:

$$D_{NM}^q(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \Big[\sum_{j=1}^{p} \phi^*(y_{i_1j}, y_{i_2j})^q\Big]^{1/q}. \tag{2.11}$$

When we want to consider the relative importance of variables, we can apply weights to Equation (2.11). This further extension is given by

$$D_{NWM}^q(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \Big[\sum_{j=1}^{p} \omega_j \big\{\phi^*(y_{i_1j}, y_{i_2j})\big\}^q\Big]^{1/q}, \tag{2.12}$$

where ω_j is a weight with $\omega_j > 0$ and $\sum_{j=1}^{p} \omega_j = 1$.

De Carvalho (1994, 1998) proposed two extensions of the Ichino-Yaguchi dissimilarity measure. In the first extension, De Carvalho introduced five comparison functions and an aggregation function such as the generalized Minkowski distance, and the second extension proposed the concept of a description potential.

Firstly, we review the first extension of the Ichino-Yaguchi measure introduced by De Carvalho (1994). The first extension uses comparison functions and an aggregation function. To define comparison functions, De Carvalho suggested agreement and disagreement indexes. For variable Y_j, the agreement and disagreement indexes between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ are summarized in Table 2.5.

Table 2.5: Agreement and disagreement indexes.

	Agreement	Disagreement	Total
Agreement	$\alpha = \lvert y_{i_1j} \cap y_{i_2j} \rvert$	$\beta = \lvert y_{i_1j} \cap c(y_{i_2j}) \rvert$	$\lvert y_{i_1j} \rvert$
Disagreement	$\chi = \lvert c(y_{i_1j}) \cap y_{i_2j} \rvert$	$\delta = \lvert c(y_{i_1j}) \cap c(y_{i_2j}) \rvert$	$\lvert c(y_{i_1j}) \rvert$
Total	$\lvert y_{i_2j} \rvert$	$\lvert c(y_{i_2j}) \rvert$	$\lvert \mathcal{Y}_j \rvert$

In Table 2.5, $|y_{ij}|$ is the length of the interval $y_{ij} = [a_{ij}, b_{ij})$, $\mathcal{Y}_j$ is the domain of variable Y_j, and $c(y_{ij})$ is the complementary set of y_{ij} in the domain. For interval-valued data, the domain $\mathcal{Y}_j$ is the interval spanned by all observations of variable Y_j, i.e., $\mathcal{Y}_j = [\min_i(a_{ij}), \max_i(b_{ij}))$. Thus, the complementary set of y_{ij} means $c(y_{ij}) = \mathcal{Y}_j - y_{ij}$.

De Carvalho proposed five comparison functions, $cf_k,\ k = 1, \ldots, 5$, using these agreement and disagreement indexes $\alpha, \beta, \chi, \delta$ defined in Table 2.5, as follows:

$$cf_1 = \frac{\alpha}{\alpha + \beta + \chi}, \tag{2.13}$$

$$cf_2 = \frac{2\alpha}{2\alpha + \beta + \chi}, \tag{2.14}$$

$$cf_3 = \frac{\alpha}{\alpha + 2(\beta + \chi)}, \tag{2.15}$$

$$cf_4 = \frac{1}{2}\left[\frac{\alpha}{\alpha + \beta} + \frac{\alpha}{\alpha + \chi}\right], \tag{2.16}$$

$$cf_5 = \frac{\alpha}{\sqrt{(\alpha + \beta)(\alpha + \chi)}}. \tag{2.17}$$

These comparison functions are defined using the similarity measures for classical binary variables and have a value between 0 and 1. Thus, since these functions have the properties of similarity measures, dissimilarity functions df_k corresponding to each comparison function are defined by

$$df_k = 1 - cf_k,\ \ k = 1, \ldots, 5. \tag{2.18}$$

Since comparison functions have a value between 0 and 1 regardless of scale of variable Y_j, comparison functions are normalized measures. Therefore, since dissimilarity functions are a linear transformation of comparison functions, they are also normalized measures.

De Carvalho's dissimilarity measure are defined using dissimilarity functions df_k, instead of ϕ or ϕ^* of the Ichino-Yaguchi measure, and the generalized Minkowski distance as an aggregation function. These dissimilarity measures between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ are given by

$$D^q_{df_k}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \left[\sum_{j=1}^{p} \left\{ \omega_j df_k(y_{i_1 j}, y_{i_2 j} \right\}^q \right]^{1/q},\ \ k = 1, \ldots, 5, \tag{2.19}$$

where ω_j is a weight with $\omega_j > 0$ and $\sum_{j=1}^{p} \omega_j = 1$.

The second extension of the Ichino-Yaguchi measure introduced by De Carvalho (1998) is defined using the description potential and an extension of Cartesian operators. Unlike the first extension, this measure does not need an aggregation function. The description potential for an interval-valued object, $\pi(\mathbf{y}_i)$, is defined by

$$\pi(\mathbf{y}_i) = \prod_{j=1}^{p} |y_{ij}|,\ \ i = 1, \ldots, n, \tag{2.20}$$

where $| \cdot |$ is a length of the interval. That is, the description potential for an interval-valued object equals the product of lengths of intervals for each variable. Also, De Carvalho defined the Cartesian operators 'join' and 'meet' between two symbolic objects. While the Cartesian operators of Equation (2.7) and (2.8) are defined for a single variable, the Cartesian operators extended by De Carvalho include all p variables. Thus, the Cartesian join between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$, $\mathbf{y}_{i_1} \oplus \mathbf{y}_{i_2}$, is defined by

$$\mathbf{y}_{i_1} \oplus \mathbf{y}_{i_2} = \left\{ y_{i_1 j} \oplus y_{i_2 j},\ \ j = 1, \ldots, p \right\}, \tag{2.21}$$

where $y_{i_1j} \oplus y_{i_2j}$ is defined in Equation (2.7). And, the Cartesian meet between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$, $\mathbf{y}_{i_1} \otimes \mathbf{y}_{i_2}$, is given by

$$\mathbf{y}_{i_1} \otimes \mathbf{y}_{i_2} = \big\{ y_{i_1j} \otimes y_{i_2j},\ j = 1, \ldots, p \big\}, \tag{2.22}$$

where $y_{i_1j} \otimes y_{i_2j}$ is defined in Equation (2.8). That is, the Cartesian join (or meet) between two symbolic objects is a set of the Cartesian join (or meet) of observations of two objects for each variable. Using these extensions, the Ichino-Yaguchi dissimilarity measure between two interval-valued objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$, $\phi_c(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$, can be extended as follows:

$$\phi_c(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \pi(\mathbf{y}_{i_1} \oplus \mathbf{y}_{i_2}) - \pi(\mathbf{y}_{i_1} \otimes \mathbf{y}_{i_2}) + \gamma\big(2\pi(\mathbf{y}_{i_1} \otimes \mathbf{y}_{i_2}) - \pi(\mathbf{y}_{i_1}) - \pi(\mathbf{y}_{i_2})\big), \tag{2.23}$$

where γ is a prespecified constant and $0 \leq \gamma \leq 0.5$, as usual. As shown in Equation (2.23), this measure does not use the aggregation function such as the Minkowski distance. De Carvalho (1998) also proposed two normalized measures of $\phi_c(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$. The first normalized measure is obtained by

$$\phi_c^*(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \frac{\phi_c(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})}{\pi(\mathcal{Y})}, \tag{2.24}$$

where $\mathcal{Y} = \big\{ \mathcal{Y}_j,\ j = 1, \ldots, p \big\}$ is the domain of variable Y_j, i.e., $\mathcal{Y}_j = \max_i(b_{ij}) - \min_i(a_{ij})$. The second normalized measure is given by

$$\phi_c'(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \frac{\phi_c(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})}{\pi(\mathbf{y}_{i_1} \oplus \mathbf{y}_{i_2})}. \tag{2.25}$$

Another distance measure for interval-valued data is the *Hausdorff distance* (Hausdorff, 1937). For variable Y_j, the Hausdorff distance between two interval-valued observations y_{i_1j} and y_{i_2j} is given by

$$\varphi(y_{i_1j}, y_{i_2j}) = \max\big[|a_{i_1j} - a_{i_2j}|, |b_{i_1j} - b_{i_2j}| \big]. \tag{2.26}$$

In the case of classical data, this measure reduces to the absolute difference between two data points. From Equation (2.26), the *Euclidean Hausdorff distance* can be defined as follows:

$$D_H(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \left[\sum_{j=1}^{p} \big\{ \varphi(y_{i_1j}, y_{i_2j}) \big\}^2 \right]^{1/2}. \tag{2.27}$$

There are two alternative normalizations of the Euclidean Hausdorff distance (see Chavent (2000) or Billard and Diday (2006)). The first normalized version can be defined as follows:

$$D_{NH_1}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \left[\sum_{j=1}^{p} \left\{ \frac{\varphi(y_{i_1 j}, y_{i_2 j})}{H_j} \right\}^2 \right]^{1/2}, \tag{2.28}$$

where

$$H_j = \frac{1}{2n^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \left\{ \varphi(y_{i_1 j}, y_{i_2 j}) \right\}^2.$$

The second version is given using the length of the domain $\mathcal{Y}_j$ as follows:

$$D_{NH_2}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}) = \left[\sum_{j=1}^{p} \left\{ \frac{\varphi(y_{i_1 j}, y_{i_2 j})}{|\mathcal{Y}_j|} \right\}^2 \right]^{1/2}, \tag{2.29}$$

where $\mathcal{Y}_j = [\min_i(a_{ij}), \max_i(b_{ij}))$ and $| \cdot |$ is the length of the interval.

In this section, we reviewed various types of similarity and dissimilarity measures for interval-valued data. Gowda and Diday (1991a,b) proposed both similarity and dissimilarity measures for symbolic objects. Both measures consist of three components representing relative size, content, and position. However, Gowda and Ravi (1995) indicated that there exists an inbalance between the Gowda-Diday similarity and dissimilarity measures, and introduced the modified version of Gowda-Diday similarity and dissimilarity measures to overcome the inbalance by using the sine and cosine functions. Ichino and Yaguchi (1994) also proposed a dissimilarity measure for symbolic objects using the Cartesian operators, and extended it to the Minkowski distance. De Carvalho (1994, 1998) suggested two extensions of the Ichino-Yaguchi dissimilarity measure. The first extension uses five comparison functions and an aggregation function such as the generalized Minkowski distance, and the second extension is defined using the description potential. In the first extension, comparison functions are defined using the agreement and disagreement indexes proposed by De Carvalho and have the types of the similarity measures for classical binary variables. In the second extension, De Carvalho defined the description potential and extended the Cartesian operators.

There exist several dis/similarity measures for histogram-valued data. Irpino and Verde (2006) proposed a distance measure for histogram data based on the Wasserstein metric, and Arroyo and Maté (2009) developed a method to forecast histogram time series using this distance measure. Also, Strelkov (2008) introduced a similarity measure focussing on the peaks of a histogram, and Cha and Srihari (2002) developed another distance measure considering correlations between ordered univariate histograms.

However, since the distance measure proposed by Irpino and Verde (2006) uses inverse functions of cumulative density functions, this measure may have some computational problems due to the invertibility of cumulative density functions. The similarity measure introduced by Strelkov (2008) accompanies complicated computations. In addition, these measures are not applicable to mixed datasets including interval-valued, multi-valued, and histogram-valued data because Irpino and Verde's measure does not provide distances for multi-valued data and Strelkov's measure cannot be extended into distances for both interval-valued and multi-valued data. The distance measure developed by Cha and Srihari (2002) deals with histograms for discrete variables. However, this study focuses on histograms for continuous variables.

Thus, to date, no dis/similarity or distance measures that are readily computable and applicable to mixed datasets exist for multivariate histogram-valued data. Therefore, in this study, we propose extended Gowda-Diday, extended Ichino-Yaguchi, and cumulative density function (CDF) dissimilarity measures for histogram-valued data, in Chapter 3.

2.4 An Overview of hierarchical divisive monothetic clustering method

One of the common issues in large dataset analysis is to detect and construct homogeneous groups from all objects in those datasets. To solve this issue, we need techniques to measure dissimilarity between objects and to classify objects. Various dissimilarity measures for interval-valued data were reviewed in Section 2.3, and in this section, we review the monothetic method that is a hierarchical divisive clustering algorithm of symbolic objects. Before reviewing the monothetic algorithm, some basics for clustering methods are explained.

Definition 2.7 *Suppose we have* p *random variables* $\{Y_j,\ j = 1, \ldots, p\}$ *with symbolic objects* $\mathbf{y}_i,\ i = 1, \ldots, n$, $\mathbf{y}_i \in \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. *Then, the* r^{th} ***partition*** *of* Ω, P_r, *is a set of subsets* $\{C_u,\ u = 1, \ldots, r\}$ *that satisfies*

$< i >$ $C_u \cap C_v = \phi$, *for all* $u \neq v = 1, \ldots, r$;

$< ii >$ $\bigcup_{u=1}^{r} C_u = \Omega$.

That is, the subsets $\{C_1, \ldots, C_r\}$ of the r^{th} partition P_r are disjoint, and exhaustive of the entire set Ω. Sometimes the subsets, $C_u,\ u = 1, \ldots, r$, are called clusters or classes.

We deal with hierarchical divisive clustering methods in this study. A hierarchy is a clustering structure.

Definition 2.8 *A* ***hierarchy*** *on* Ω *is a set of subsets* $H = \{C_u,\ u = 1, \ldots, r\}$ *that satisfies the following properties:*

(i) $\Omega \in H$;

(ii) for all single objects $\mathbf{y}_i$ *in* Ω, $\{\mathbf{y}_i\} \in H$;

(iii) for all $C_u,\ C_v \in H,\ u \neq v = 1, \ldots, r$, $C_u \cap C_v \in \{\phi, C_u, C_v\}$.

That is, property (iii) means that either any two clusters C_u and C_v are disjoint, or one cluster is contained in the other one. A hierarchical clustering consists of a series of partitions. The clusters of a hierarchical classification can be displayed by a rooted tree where the root is the entire set Ω and n clusters with a single object as the leaves of the tree. This tree structured plot is called a dendrogram.

A hierarchical clustering is typically performed in a recursive way that either goes ‘from top to bottom’ by successive splitting of clusters or ‘from bottom to top’ by successive agglomeration of clusters. The former is called the *divisive clustering method*, and the latter is called the *agglomerative clustering method*. In general, the agglomerative clustering method starts with n clusters each with a single object, and successively merges two clusters using some clustering criteria. Finally, it reaches the entire set Ω. In contrast, the divisive clustering method starts from a single cluster containing all objects in Ω. At each stage, a cluster is partitioned into two clusters. This is repeated until all clusters have only one object.

In this study, we focus on the divisive clustering method. In hierarchical clustering methods, the agglomerative clustering method is more widely used than the divisive clustering method because the divisive clustering has too many possible bipartitions $(2^{n-1} - 1)$ and computationally spends much more time and costs. However, the divisive clustering has some advantages. It shows the main structure in datasets and avoids unfortunate decisions at earlier stages. In addition, the monothetic and polythetic algorithms to be introduced in this section and Chapter 3, respectively, can reduce the number of possible bipartitions to be examined.

Clustering criteria are also important in cluster analysis because they measure the quality of a partition and affect outcomes of clustering. That is, different clustering criteria can lead to different clustering outcomes. There are various clustering criteria such as minimum distance (or, single linkage), maximum distance (complete linkage), average distance, and Ward criterion, etc. One well-known criterion is the variance criterion. In the variance criterion, the optimal partition has the minimum variance.

In this study, we focus on divisive clustering methods of symbolic objects. There are various divisive clustering methods for classical data. Edward and Cavalli-Sforza (1965) introduced a divisive clustering method that finds the optimal bipartition among the $2^{n-1}-1$ possible bipartitions for a cluster with n objects using the within-cluster sum of squares as a clustering criterion. Since this method considers all possible bipartitions in a cluster, it always gives a global optimal bipartition but is not computationally efficient. MacNaughton-Smith et al. (1964) proposed a method that iteratively uses an average distance between an object and a group of objects. Also, Har-even and Brailovsky (1995) introduced a probabilistic validation approach for divisive clustering.

Williams and Lambert (1959) and Lance and Williams (1968) have first proposed monothetic divisive clustering methods for classical binary data. For symbolic data, Chavent (1998, 2000) developed a divisive clustering method for interval data using a hierarchy of a set of objects and a monothetic characteristic of each cluster of the hierarchy. At each step, bipartitioning is performed by minimizing the within-cluster variance, and a binary question corresponding to a monothetic characteristic can be found. A brief description of Chavent's method follows.

Suppose that there are p interval-valued random variables $\{Y_j,\ j = 1, \ldots, p\}$ with observations $\mathbf{y}_i = \{y_{i1}, \ldots, y_{ip}\} \in \Omega$ for $i = 1, \ldots, n$, and let P_r be a r^{th} stage partition. Then, a partition of Ω at the r^{th} stage, P_r, can be represented by a set of subsets $\{C_u,\ u = 1, \ldots, r\}$. The subset, C_u, is called a cluster. At the $(r+1)^{th}$ stage a single cluster C_u in the partition P_r is bisected into C_u^1 and C_u^2. Thus, a new partition can be written as

$$P_{r+1} = \big(P_r \cup \{C_u^1, C_u^2\}\big) - \{C_u\}.$$

To perform divisive hierarchical clustering, Chavent (1998, 2000) proposed the within-cluster variance as a criterion partitioning a cluster. For a cluster $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$, the *within-cluster variance* $I(C_u)$ is defined by

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}), \tag{2.30}$$

where $D(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ is a dissimilarity or distance measure between the objects $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ in C_u and w_i is the weight for the $\mathbf{y}_i$, and where $\tau = \sum_{i=1}^{n_u} w_i$. The $I(C_u)$ is a homogeneity measure for the cluster C_u. Often, it is desired to weight each object according to its size. In that case, w_i can be the proportion of the size of an object to the total size over all objects. However, if the weight for each observation is equivalent, $w_i = 1/n$, where n is the total number of objects in Ω ($n = \sum_{u=1}^{r} n_u$). The *total within-cluster variance* for a partition P_r is the sum of within-cluster variances for all clusters in P_r. It is written as

$$W(P_r) = \sum_{u=1}^{r} I(C_u). \tag{2.31}$$

The *between-cluster variance* for the partition P_r is defined by

$$B(P_r) = W(\Omega) - W(P_r), \tag{2.32}$$

where $\Omega \equiv P_1$. The P_1 means there is only one cluster in the partition and includes all objects. From (2.32), we know that minimizing the within-cluster variance is equivalent to maximizing the between-cluster variance.

A characteristic of the divisive monothetic clustering is that the partition can be bisected by a binary question for a single variable at each stage because it considers bipartitions sorted by each variable to find a bipartition (C_u^1, C_u^2) minimizing the total within-cluster variance. The form of a binary question is 'Is $Y_j \leq c$?', where c is the *cut point*. For the interval-valued variable Y_j taking values $[a_{ij}, b_{ij}),\ i = 1, \dots, n,\ j = 1, \dots, p$, suppose that the cluster $C_u = \{\mathbf{y}_1, \dots, \mathbf{y}_{n_u}\}$ is bisected into C_u^1 and C_u^2. Then, candidates c_{qj} for the cut point c for a variable Y_j are obtained by

$$c_{qj} = (\bar{y}_{qj} + \bar{y}_{q+1,j})/2,\ q = 1, \dots, n_u - 1, \tag{2.33}$$

where $\bar{y}_{qj}$ is the value sorted by ascending values of $\bar{y}_{ij}$, where $\bar{y}_{ij} = (a_{ij} + b_{ij})/2$. Thus, there exist $p(n_u - 1)$ candidates for the cut point in the cluster C_u. This means that the (C_u^1, C_u^2) minimizing the total within-cluster variance can be chosen among the $p(n_u - 1)$ bipartitions corresponding to each candidate for cut point, and the cut point c is the c_{qj} corresponding to the selected (C_u^1, C_u^2). Therefore, each object is classified into C_u^1 or C_u^2 by whether the answer for the detected binary question 'Is $Y_j \leq c$?' is 'yes' or 'no'.

Table 2.6: Interval-valued data for Ruspini data.

$\mathbf{y}$	Y_1	Y_2
$\mathbf{y}_1$	[4,36]	[49,88]
$\mathbf{y}_2$	[28,63]	[124,156]
$\mathbf{y}_3$	[74,117]	[94,132]
$\mathbf{y}_4$	[58,83]	[4,31]

To understand the monothetic algorithm, we see an example for the Ruspini (1970) data well known in cluster analysis. The classical Ruspini dataset is artificial data and consists of

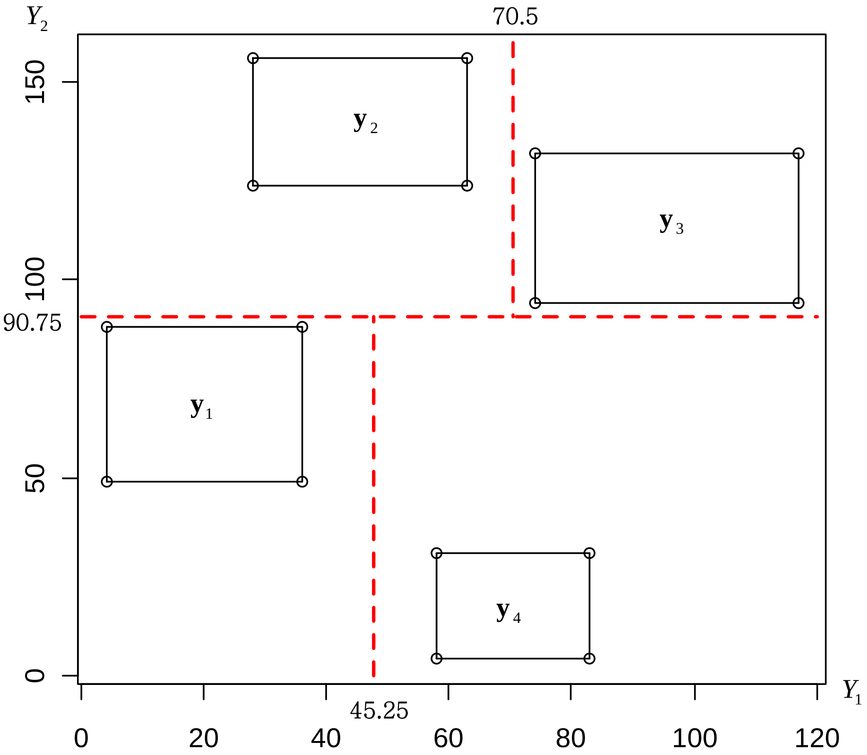


Figure 2.1: 4 interval-valued objects for Ruspini data.

75 observations with two variables; and it is also well known that there are four clusters in this dataset. In our example, we use these four clusters as interval-valued objects to explain the monothetic algorithm. Thus, we have four interval-valued objects (i.e., $\mathbf{y}_1, \ldots, \mathbf{y}_4$) with two variables (i.e., Y_1 and Y_2) as shown in Table 2.6, and the plot for these interval-valued data, represented by rectangles, is shown in Figure 2.1.

Since this is a divisive clustering method, we start with all four objects. In addition, since there are four objects ($n = 4$), the number of all possible biparitions is seven ($= 2^{4-1} - 1$) at the first stage. However, seven possible bipartitions can be reduced to six possible bipartitions ($= 2(4-1)$) by the monothetic method. In this example, the number of possible bipartitions to be reduced by the monothetic method is small becasue the number of objects is also small. In general, that number would be much larger for a large number of objects.

Table 2.7: $\bar{y}_{ij}$ values for Ruspini interval-valued data.

$\bar{y}_{ij}$	Y_1	Y_2
$\mathbf{y}_1$	$\bar{y}_{11} = 20$	$\bar{y}_{12} = 68.5$
$\mathbf{y}_2$	$\bar{y}_{21} = 45.5$	$\bar{y}_{22} = 140$
$\mathbf{y}_3$	$\bar{y}_{31} = 95.5$	$\bar{y}_{32} = 113$
$\mathbf{y}_4$	$\bar{y}_{41} = 70.5$	$\bar{y}_{42} = 17.5$

Table 2.7 shows the mid-point values for intervals, $\bar{y}_{ij} = (a_{ij}+b_{ij})/2,\ i = 1, \ldots, 4,\ j = 1, 2$. From this table, ascending orders of objects for each variable can be obtained by ascending values of $\bar{y}_{ij}$. For the variable Y_1, $\mathbf{y}_1 < \mathbf{y}_2 < \mathbf{y}_4 < \mathbf{y}_3$, and for the variable Y_2, $\mathbf{y}_4 < \mathbf{y}_1 < \mathbf{y}_3 < \mathbf{y}_2$. Using these orders, possible bipartitions for the variable Y_1 are $\big\{(\mathbf{y}_1), (\mathbf{y}_2, \mathbf{y}_4, \mathbf{y}_3)\big\}$, $\big\{(\mathbf{y}_1, \mathbf{y}_2), (\mathbf{y}_4, \mathbf{y}_3)\big\}$, $\big\{(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_4), (\mathbf{y}_3)\big\}$; and for the variable Y_2, we have $\big\{(\mathbf{y}_4), (\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_2)\big\}$, $\big\{(\mathbf{y}_4, \mathbf{y}_1), (\mathbf{y}_3, \mathbf{y}_2)\big\}$, $\big\{(\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_3), (\mathbf{y}_2)\big\}$.

Suppose that the bipartition $\big\{(\mathbf{y}_4, \mathbf{y}_1), (\mathbf{y}_3, \mathbf{y}_2)\big\}$ made by the variable Y_2 has the minimum value of the sum of within-cluster variance in Equation (2.31). Then, the cut point at the first stage can be obtained by

$$c = (\bar{y}_{12} + \bar{y}_{32})/2 = (68.5 + 113)/2 = 90.75,$$

and the binary question is 'Is $Y_2 \leq 90.75$?'. If for a given object the answer for this binary question is 'Yes', that object goes to the cluster $\big\{\mathbf{y}_4, \mathbf{y}_1\big\}$, and if 'No', it goes to the cluster $\big\{\mathbf{y}_3, \mathbf{y}_2\big\}$.

Let cluster $C_1 = \big\{\mathbf{y}_4, \mathbf{y}_1\big\}$ and $C_2 = \big\{\mathbf{y}_3, \mathbf{y}_2\big\}$. Then, either C_1 or C_2 is bipartitioned at the second stage. For the variable Y_1, ascending orders for each cluster are C_1; $\mathbf{y}_1 < \mathbf{y}_4$ and C_2; $\mathbf{y}_2 < \mathbf{y}_3$, and for the variable Y_2, C_1; $\mathbf{y}_4 < \mathbf{y}_1$ and C_2; $\mathbf{y}_3 < \mathbf{y}_2$. Although these ascending orders are meaningless (in the present example) because there are only two objects in each cluster, this procedure is necessary if there are more than two objects in a cluster. Thus, in this case, we have less than four possible bipartitions $(= 2(2-1) + 2(2-1))$ from the monothetic method, and the actual number of possible bipartitions is

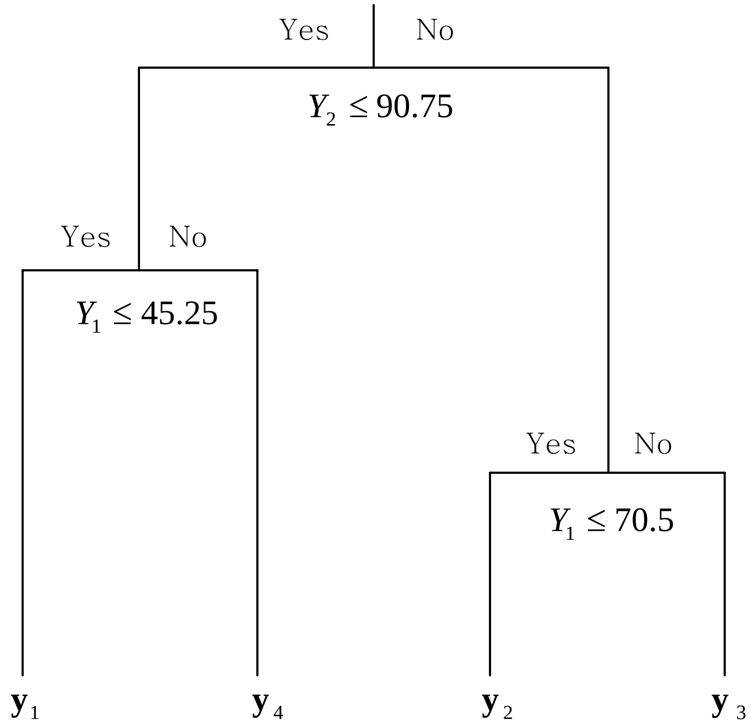


Figure 2.2: Dendrogram for Ruspini interval-valued data.

two, $\{(\mathbf{y}_1),(\mathbf{y}_4),(\mathbf{y}_2,\mathbf{y}_3)\}$ and $\{(\mathbf{y}_1,\mathbf{y}_4),(\mathbf{y}_2),(\mathbf{y}_3)\}$. Suppose that the optimal bipartition is $\{(\mathbf{y}_1),(\mathbf{y}_4),(\mathbf{y}_2,\mathbf{y}_3)\}$, i.e., C_1 is bipartitioned into $C_1^1 = \{\mathbf{y}_1\}$ and $C_1^2 = \{\mathbf{y}_4\}$. Then, there are two possible binary questions because the orders for variables Y_1 and Y_2 are the same. For the variable Y_1,

$$c = (\bar{y}_{11} + \bar{y}_{41})/2 = (20 + 70.5)/2 = 45.25,$$

and for the variable Y_2,

$$c = (\bar{y}_{12} + \bar{y}_{42})/2 = (68.5 + 17.5)/2 = 43.$$

Thus, there are two possible binary questions, 'Is $Y_1 \leq 45.25$?' and 'Is $Y_2 \leq 43$?', but we have to choose one of them by using the dissimilarity or distance measure values between $\mathbf{y}_1$ and $\mathbf{y}_4$ for each variable. That is, we compare the distance between two observations y_{11} and

y_{41}, $D(y_{11}, y_{41})$ with the distance between y_{12} and y_{42}, $D(y_{12}, y_{42})$. Suppose that $D(y_{11}, y_{41})$ is larger than $D(y_{12}, y_{42})$; then the binary question of the second stage is 'Is $Y_1 \leq 45.25$?'.

Since the third stage is the last stage, $\{\mathbf{y}_2, \mathbf{y}_3\}$ should be bipartitioned into $\{\mathbf{y}_2\}$ and $\{\mathbf{y}_3\}$. Similarly to the second stage, suppose the distance $D(y_{21}, y_{31})$ is larger than $D(y_{22}, y_{32})$; then the binary question of the third stage is 'Is $Y_1 \leq 70.5$?'. These results are shown using the dendrogram in Figure 2.2. From Figure 2.2, if the answer for the binary question of the first stage is 'Yes', then we ask 'Is $Y_1 \leq 45.25$?'. If 'No', then we ask 'Is $Y_1 \leq 70.5$?'.

In summary, Chavent's monothetic method uses the within-cluster variance as a clustering criterion. This monothetic method reduces the amount of calculation to detect the optimal (C_u^1, C_u^2) in the C_u. That is, suppose that we want to find the optimal bipartition at the $(r+1)^{th}$ stage from the partition $P_r = \{C_u,\ u = 1, \ldots, r\}$; then there exist a total of $\sum_{u=1}^{r}(2^{n_u-1} - 1)$ possible bipartitions. However, the monothetic method needs to examine only $\sum_{u=1}^{r} p(n_u - 1)$ bipartitions to find the optimal bipartition. When the number of objects is large, the number of possible bipartitions is even further reduced by the monothetic method. However, since the monothetic method is based on a single variable to detect the optimal bipartition at each stage, it performs poorly in a structure that depends on combinations of variables. Our new proposed polythetic algorithm (in Section 3.4.2) overcomes this deficiency.

Chapter 3

Clustering for Histogram-valued Data

In this chapter, we focus on dissimilarity measures and divisive clustering methods for histogram-valued data. In Section 3.1, the histogram-valued random variable is defined; and a transformation for histogram-valued data to enable the new methodologies to be handled more easily computationally is proposed in Section 3.2. To perform clustering, basically we need a dissimilarity/distance measure, a clustering criterion, and an algorithm. Thus, three dissimilarity measures using transformed histogram-valued data are introduced in Section 3.3. In Section 3.4, we show how to extend the monothetic method to histogram-valued observations. A polythetic clustering method based on all p variables is proposed in Section 3.4.

3.1 Histogram-valued Data

A histogram is a useful tool to summarize data graphically. It includes information for the approximate shape and location for data and is specified by non-overlapping subintervals and their frequency. Thus, a histogram-valued random variable can be defined as follows.

Definition 3.1 *Let Y be a random variable that takes values on a finite number of non-overlapping intervals $\{[a_k, a_{k+1}),\ k = 1, \ldots, v\}$ with relative frequencies p_k corresponding to each subinterval, where $a_k \leq a_{k+1}$. Then, Y is called a* ***histogram-valued random variable****. The i^{th} observation y_i for a histogram-valued random variable is given by*

$$y_i = \Big\{[a_{ik}, a_{i,k+1}),\ p_{ik};\ k = 1, \ldots, v_i\Big\}, \tag{3.1}$$

where $\sum_{k=1}^{v_i} p_{ik} = 1,\ i = 1, \ldots, n.$

Definition 3.2 *Let* $\mathbf{Y} = (Y_1, \ldots, Y_p)$ *be a* p*-dimensional histogram-valued random variable. Then, the* i^{th} *observation* $\mathbf{y}_i$ *can be written in the form, for each* $i = 1, \ldots, n$*,*

$$\mathbf{y}_i = \big([a_{ijk}, a_{ij,k+1}),\ p_{ijk};\ j = 1, \ldots, p,\ k = 1, \ldots, v_{ij}\big), \tag{3.2}$$

where $\sum_{k=1}^{v_{ij}} p_{ijk} = 1$.

3.2 Data Transformation

Unlike interval-valued data, it is not easy to deal with histogram-valued data computationally nor to apply existing dissimilarity measures to them because each observation has different lengths and numbers of subintervals. If we transform them so as to have the same length and number of subintervals for each variable, handling such data in a computer would be easier and existing dissimilarity measures could be extended to histogram-valued data. Thus, we try to transform the observations into a form which has the same length and number of subintervals. Once transformed subintervals are set up, then relative frequencies can be calculated for the new subintervals.

Definition 3.3 *Let* $\{[b_{jk}, b_{j,k+1}),\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\}$ *be a transformed subinterval for the* j^{th} *variable. Then, we can define*

$$b_{j1} = \min_i\{a_{ij1}\}, \tag{3.3}$$

$$b_{j,t_j+1} = \max_i\{a_{ij,t_{ij}+1}\}, \tag{3.4}$$

and

$$b_{j,k+1} = b_{j1} + k\Psi_j/t_j,\ k = 1, \ldots, t_j, \tag{3.5}$$

where

$$\Psi_j = b_{j,t_j+1} - b_{j1}, \tag{3.6}$$

$$t_j = \left\lceil \frac{\Psi_j}{\min_{i,k}\{a_{ij,k+1} - a_{ijk}\}} \right\rceil, \tag{3.7}$$

where $\lceil\cdot\rfloor$ is rounding off a number to the nearest integer. Thus, a transformed histogram-valued observation can be written as

$$\mathbf{y}'_i = \{y'_{ij},\ j=1,\ldots,p\} = \{[b_{jk}, b_{j,k+1}),\ p'_{ijk};\ j=1,\ldots,p,\ k=1,\ldots,t_j\},\ i=1,\ldots,n, \tag{3.8}$$

where p'_{ijk} is a transformed relative frequency corresponding to a transformed subinterval, and $\sum_{k=1}^{t_j} p'_{ijk} = 1$.

This subinterval does not depend on any one observation. That is, each observation has the same subintervals (i.e., same subinterval and points) and the same number of subintervals for each variable; but can differ for different variables. Only relative frequency values are different for the observations. The lengths of the transformed subinterval for each variable are the same (i.e., $b_{j,k+1} - b_{jk} = b_{j,k'+1} - b_{jk'}$, for all $k,\ k' = 1, \ldots, t_j$). In addition, the length of the transformed subinterval for each variable is approximate or equal to the minimum length of the original subintervals as shown in Equation (3.7).

As the length of a subinterval increases, the number of subintervals decreases but the loss of information increases. On the contrary, if the length of a subinterval is small, we can minimize the loss of information for the original histogram-valued data. However, if most objects for a variable Y_j have similar lengths of subintervals and only a few objects have relatively very small lengths, the number of transformed subintervals would be very large. This has a consequence that calculations become computationally very expensive. In that case, we might consider the average or mode of lengths of the original subintervals, instead of the minimum length.

For example, consider histogram-valued data for a variable Y_1 as shown in Table 3.1. The lengths of subintervals for all objects have a value between 0.1 and 4. In contrast, the lengths of subintervals except for $\mathbf{y}_5$ have a value between 1 and 4. Suppose the object $\mathbf{y}_5$ is excluded in the dataset. That is, the dataset has histogram-valued objects $\mathbf{y}_1, \ldots, \mathbf{y}_4$. Then, using Definition 3.3,

Table 3.1: An example for histogram-valued data with different subintervals.

	Y_1
$\mathbf{y}_1$	{[4,6),0.08; [6,10),0.24; [10,12),0.30; [12,18),0.30; [18,20),0.08}
$\mathbf{y}_2$	{[6,8),0.08; [8,9),0.40; [9,12),0.24; [12,14),0.24; [14,15),0.04}
$\mathbf{y}_3$	{[1,3),0.18; [3,5),0.20; [5,8),0.42; [8,10),0.14; [10,14),0.06}
$\mathbf{y}_4$	{[2,4),0.18; [4,8),0.60; [8,12),0.16; [12,15),0.06}
$\mathbf{y}_5$	{[0,0.1),0.10; [0.1,0.3),0.20; [0.3,0.6),0.26; [0.6,0.8),0.28; [0.8,1.0),0.16}

$$b_{11} = \min_i\{a_{i11}\} = \min\{4, 6, 1, 2\} = 1,$$

and

$$b_{1,t_1+1} = \max_i\{a_{i1,t_{i1}+1}\} = \max\{20, 15, 14, 15\} = 20.$$

Thus,

$$\Psi_1 = b_{1,t_1+1} - b_{11} = 20 - 1 = 19,$$

$$t_1 = \left\lceil \frac{\Psi_1}{\min_{i,k}\{a_{i1,k+1} - a_{i1k}\}} \right\rceil = \left\lceil \frac{19}{1} \right\rceil = 19,$$

and

$$b_{1,k+1} = b_{11} + k\Psi_1/t_1 = 1 + k,\ \ k = 1, \ldots, 19.$$

Thus, in this case we have 19 transformed subintervals with length 1. In contrast, if the dataset includes the object $\mathbf{y}_5$, then

$$b_{11} = \min_i\{a_{i11}\} = \min\{4, 6, 1, 2, 0\} = 0,$$

and

$$b_{1,t_1+1} = \max_i\{a_{i1,t_{i1}+1}\} = \max\{20, 15, 14, 15, 1\} = 20.$$

Thus,

$$\Psi_1 = b_{1,t_1+1} - b_{11} = 20 - 0 = 20,$$

$$t_1 = \left\lceil \frac{\Psi_1}{\min_{i,k}\{a_{i1,k+1} - a_{i1k}\}} \right\rceil = \left\lceil \frac{20}{0.1} \right\rceil = 200,$$

and

$$b_{1,k+1} = b_{11} + k\Psi_1/t_1 = k(0.1), \;\; k = 1, \ldots, 200.$$

In this case, we have 200 transformed subintervals with length 0.1. Since the minimum length of $\mathbf{y}_5$ is relatively very small to other objects, the number of subintervals for including $\mathbf{y}_5$ is 10 times more than that for excluding $\mathbf{y}_5$. This has a consequence that the cost for computing steeply increases. Thus, in this case, it might be better to consider the average or mode of lengths of the original subintervals, rather than the minimum length.

Now, we should calculate the transformed relative frequencies for each observation because the subintervals are changed. The relative frequencies for the transformed subintervals are calculated from the overlapping proportion between the original and the new subintervals. The relative frequency for any non-overlapped portion is assigned to zero. This is illustrated through the following example.

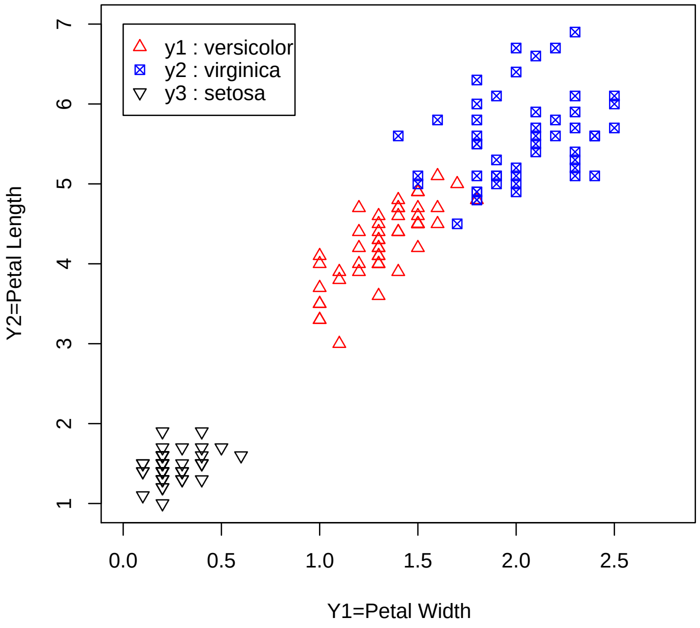


Figure 3.1: Two-dimensional plot for the iris data

Table 3.2: Petal width and petal length for iris species.

$\mathbf{y}$	Species	Y_1 =Petal Width	Y_2 =Petal Length
$\mathbf{y}_1$	versicolor	$\{[1.0, 1.5), 0.90;\ [1.5, 2.0), 0.10\}$	$\{[3.0, 4.0), 0.32;\ [4.0, 5.0), 0.66;\ [5.0, 6.0), 0.02\}$
$\mathbf{y}_2$	virginica	$\{[1.2, 1.6), 0.08;\ [1.6, 2.0), 0.46;\ [2.0, 2.4), 0.40;\ [2.4, 2.8), 0.06\}$	$\{[4.5, 5.5), 0.50;\ [5.5, 6.5), 0.42;\ [6.5, 7.5), 0.08\}$
$\mathbf{y}_3$	setosa	$\{[0.0, 0.4), 0.96;\ [0.4, 0.8), 0.04\}$	$\{[1.0, 1.5), 0.74;\ [1.5, 2.0), 0.26\}$

Example 3.1 *Consider the three sets of observations from Fisher's (1936) iris data shown in Table 3.2 obtained by aggregating observations by species. This gives histogram values for Y_1 ='Petal Width' and Y_2 ='Petal Length' for the three species of iris, iris versicolor (object $\mathbf{y}_1$), virginica ($\mathbf{y}_2$), and setosa ($\mathbf{y}_3$). Figure 3.1 shows the classical data points for the complete dataset of 150 observations. Now, we calculate the transformed subinterval and relative frequency. From Equation (3.3), (3.4), (3.5) and (3.7), the transformed subinterval for Y_1 is given by*

$$b_{11} = \min\{1, 1.2, 0\} = 0, \quad b_{1,t_1+1} = \max\{2, 2.8, 0.8\} = 2.8,$$

$$t_1 = \left\lceil \frac{2.8 - 0}{\min\{0.5, 0.4, 0.4\}} \right\rfloor = 7,$$

and

$$b_{12} = 0 + 1\left(\frac{2.8 - 0}{7}\right) = 0.4, \quad b_{13} = 0 + 2\left(\frac{2.8}{7}\right) = 0.8, \ldots, \quad b_{17} = 0 + 6\left(\frac{2.8}{7}\right) = 2.4.$$

Thus, the transformed subinterval for Y_1 is $\{[0, 0.4),\ [0.4, 0.8), \ldots,\ [2.4, 2.8)\}$. For the observation y_{11}, the relative frequency corresponding to the transformed subinterval is obtained by

$$p'_{111} = 0, \quad p'_{112} = 0, \quad p'_{113} = 0.9\left(\frac{1.2 - 1}{1.5 - 1}\right) = 0.36,$$
$$p'_{114} = 0.9\left(\frac{1.5 - 1.2}{1.5 - 1}\right) + 0.1\left(\frac{1.6 - 1.5}{2 - 1.5}\right) = 0.56,$$
$$p'_{115} = 0.1\left(\frac{2 - 1.6}{2 - 1.5}\right) = 0.08, \quad p'_{116} = 0, \quad p'_{117} = 0.$$

Table 3.3: The transformed histogram values for the data of Table 3.2.

$\mathbf{y}'$	Species	Y_1 =Petal Width	Y_2 =Petal Length
$\mathbf{y}'_1$	versicolor	$\{[0.0, 0.4), 0.00;\ [0.4, 0.8), 0.00;$ $[0.8, 1.2), 0.36;\ [1.2, 1.6), 0.56;$ $[1.6, 2.0), 0.08;\ [2.0, 2.4), 0.00;$ $[2.4, 2.8), 0.00\}$	$\{[1.0, 1.5), 0.00;\ [1.5, 2.0), 0.00;$ $[2.0, 2.5), 0.00;\ [2.5, 3.0), 0.00;$ $[3.0, 3.5), 0.16;\ [3.5, 4.0), 0.16;$ $[4.0, 4.5), 0.33;\ [4.5, 5.0), 0.33;$ $[5.0, 5.5), 0.01;\ [5.5, 6.0), 0.01;$ $[6.0, 6.5), 0.00;\ [6.5, 7.0), 0.00;$ $[7.0, 7.5), 0.00\}$
$\mathbf{y}'_2$	virginica	$\{[0.0, 0.4), 0.00;\ [0.4, 0.8), 0.00;$ $[0.8, 1.2), 0.00;\ [1.2, 1.6), 0.08;$ $[1.6, 2.0), 0.46;\ [2.0, 2.4), 0.40;$ $[2.4, 2.8), 0.06\}$	$\{[1.0, 1.5), 0.00;\ [1.5, 2.0), 0.00;$ $[2.0, 2.5), 0.00;\ [2.5, 3.0), 0.00;$ $[3.0, 3.5), 0.00;\ [3.5, 4.0), 0.00;$ $[4.0, 4.5), 0.00;\ [4.5, 5.0), 0.25;$ $[5.0, 5.5), 0.25;\ [5.5, 6.0), 0.21;$ $[6.0, 6.5), 0.21;\ [6.5, 7.0), 0.04;$ $[7.0, 7.5), 0.04\}$
$\mathbf{y}'_3$	setosa	$\{[0.0, 0.4), 0.96;\ [0.4, 0.8), 0.04;$ $[0.8, 1.2), 0.00;\ [1.2, 1.6), 0.00;$ $[1.6, 2.0), 0.00;\ [2.0, 2.4), 0.00;$ $[2.4, 2.8), 0.00\}$	$\{[1.0, 1.5), 0.74;\ [1.5, 2.0), 0.26;$ $[2.0, 2.5), 0.00;\ [2.5, 3.0), 0.00;$ $[3.0, 3.5), 0.00;\ [3.5, 4.0), 0.00;$ $[4.0, 4.5), 0.00;\ [4.5, 5.0), 0.00;$ $[5.0, 5.5), 0.00;\ [5.5, 6.0), 0.00;$ $[6.0, 6.5), 0.00;\ [6.5, 7.0), 0.00;$ $[7.0, 7.5), 0.00\}$

Thus, the transformed histogram value for y_{11} *is* $\{[0.0, 0.4), 0.00;\ [0.4, 0.8), 0.00;\ [0.8, 1.2), 0.36;\ [1.2, 1.6), 0.56;\ [1.6, 2.0), 0.08;\ [2.0, 2.4), 0.00\ [2.4, 2.8), 0.00\}$. *Note, the sum of transformed relative frequencies is 1 (*$= \sum_{k=1}^{t_j} p'_{ijk}$*). Figure 3.2 illustrates the procedure to obtain* p'_{11k}, $k = 1, \ldots, 7$. *Similarly, we can transform all the data of Table 3.2. This completed result is shown in Table 3.3 for both* Y_1 *and* Y_2 *and all three observations.*

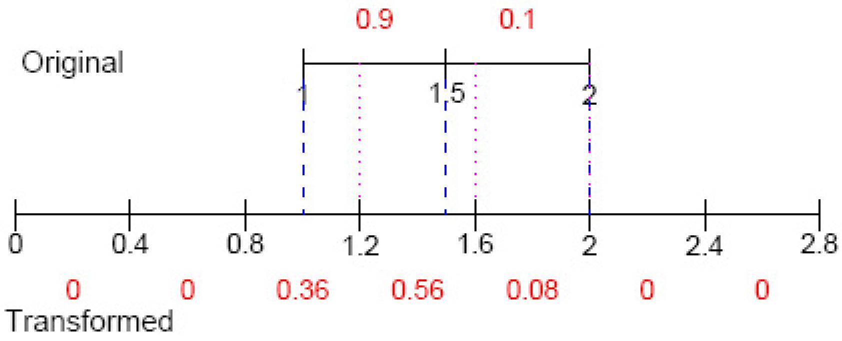


Figure 3.2: The transformed relative frequencies for y_{11}.

3.3 Dissimilarity and Distance Measures

In this section, we introduce dissimilarity and distance measures for histogram-valued data. Usually we judge how far an object is from another object using these measures. Since basically most clustering methods depend on them, it is very important to measure dissimilarity or distance accurately and reasonably.

For continuous variables, the dissimilarity for classical data is interpreted as the distance between two objects based on their location in p-dimensional space. In contrast, even when the centers of two symbolic objects are located in the same point, we cannot say they are similar because the degree of their dispersion may be different. Thus, the dissimilarity for histogram-valued data should reflect both their location and dispersion. We first propose dissimilarity or distance measures for histogram-valued data in this section by extending two particular measures developed for the case of interval-valued data by Gowda and Diday (1991a) and Ichino and Yaguchi (1994). There are followed by a measure based on the cumulative density function for histogram data (see Definition 3.12).

In order to define dissimilarity measures, first of all, we define the union and intersection between two histogram-valued objects, and the mean and standard deviation for histogram-valued data.

Definition 3.4 *Let* $\mathbf{y}'_i$ *be a transformed histogram-valued observation, corresponding to* $\mathbf{y}_i$*, with subintervals and relative frequencies* $\left([b_{jk}, b_{j,k+1}),\ p'_{ijk};\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\right),\ i = 1, \ldots, n$*; and let* $\mathbf{y}'_{(i_1 \cup i_2)}$ *be the* ***union*** *between two transformed histogram-valued observations* $\mathbf{y}'_{i_1}$ *and* $\mathbf{y}'_{i_2}$*. Then,* $\mathbf{y}'_{(i_1 \cup i_2)}$ *is*

$$\mathbf{y}'_{(i_1 \cup i_2)} = \{[b_{jk}, b_{j,k+1}),\ p'_{(i_1 \cup i_2)jk};\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\},\ i_1, i_2 = 1, \ldots, n, \tag{3.9}$$

where

$$p'_{(i_1 \cup i_2)jk} = \max\{p'_{i_1jk}, p'_{i_2jk}\},\ k = 1, \ldots, t_j. \tag{3.10}$$

Also, let $\mathbf{y}'_{(i_1 \cap i_2)}$ *be the* ***intersection*** *between two histogram-valued observations* $\mathbf{y}'_{i_1}$ *and* $\mathbf{y}'_{i_2}$*. Then,* $\mathbf{y}'_{(i_1 \cap i_2)}$ *is given by*

$$\mathbf{y}'_{(i_1 \cap i_2)} = \{[b_{jk}, b_{j,k+1}),\ p'_{(i_1 \cap i_2)jk};\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\},\ i_1, i_2 = 1, \ldots, n, \tag{3.11}$$

where

$$p'_{(i_1 \cap i_2)jk} = \min\{p'_{i_1jk}, p'_{i_2jk}\},\ k = 1, \ldots, t_j. \tag{3.12}$$

Note that in Equations (3.10) and (3.12), $\sum_{k=1}^{t_j} p'_{(i_1 \cup i_2)jk} \geq 1$ and $\sum_{k=1}^{t_j} p'_{(i_1 \cap i_2)jk} \leq 1$, respectively.

Figure 3.3 displays an example of the union and intersection, respectively, of two histogram-valued observations for a variable Y_j.

From Billard and Diday (2003), we have the descriptive statistics, empirical mean and variance for histogram-valued data, as follows:

Definition 3.5 *Let* y'_{ij} *be a transformed histogram-valued observation for a variable* Y_j*. Then, the mean for* y'_{ij} *is defined by*

$$M_{ij} = \sum_{k=1}^{t_j} \left(\frac{b_{jk} + b_{j,k+1}}{2}\right) p'_{ijk}, \tag{3.13}$$

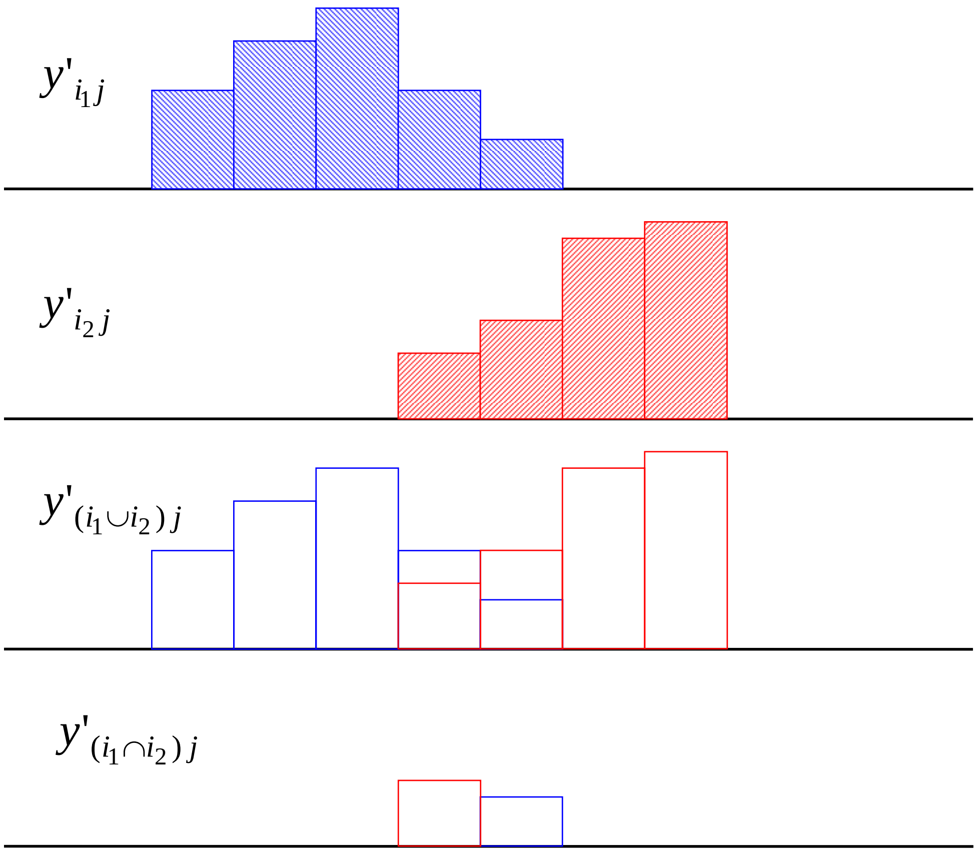


Figure 3.3: An example of the union and intersection between two histogram-valued data.

and the standard deviation for y'_{ij} is given by

$$S_{ij} = \sqrt{\sum_{k=1}^{t_j} \left\{ \frac{(b_{jk} - M_{ij})^2 + (b_{jk} - M_{ij})(b_{j,k+1} - M_{ij}) + (b_{j,k+1} - M_{ij})^2}{3} \right\} p'_{ijk}} \, . \qquad (3.14)$$

However, the means for the union and intersection between two objects are different from that of Definition 3.5. For two single observations $y'_{i_1 j}$ and $y'_{i_2 j}$, if we use $p'_{(i_1 \cup i_2) jk}$ or $p'_{(i_1 \cap i_2) jk}$ to obtain the mean, it is not a measure of the mean because $\sum_{k=1}^{t_j} p'_{(i_1 \cup i_2) jk} \geq 1$ and $\sum_{k=1}^{t_j} p'_{(i_1 \cap i_2) jk} \leq 1$. Thus, the sum of $p'_{(i_1 \cup i_2) jk}$ and $p'_{(i_1 \cap i_2) jk}$, respectively, need to be standardized to 1. This leads to the following definitions for the empirical mean and variance of the union and intersection of histogram-valued data.

Definition 3.6 *Let $y'_{(i_1 \cup i_2)j}$ be the union between two transformed histogram-valued observations y'_{i_1j} and y'_{i_2j}. Then, the mean of their union $y'_{(i_1 \cup i_2)j}$ is*

$$M^*_{(i_1 \cup i_2)j} = \sum_{k=1}^{t_j} \left(\frac{b_{jk} + b_{j,k+1}}{2} \right) p^*_{(i_1 \cup i_2)jk}, \tag{3.15}$$

where

$$p^*_{(i_1 \cup i_2)jk} = \frac{p'_{(i_1 \cup i_2)jk}}{\sum_{k=1}^{t_j} p'_{(i_1 \cup i_2)jk}}. \tag{3.16}$$

Also, the mean for the intersection $y'_{(i_1 \cap i_2)j}$ can be written as

$$M^*_{(i_1 \cap i_2)j} = \sum_{k=1}^{t_j} \left(\frac{b_{jk} + b_{j,k+1}}{2} \right) p^*_{(i_1 \cap i_2)jk}, \tag{3.17}$$

where

$$p^*_{(i_1 \cap i_2)jk} = \frac{p'_{(i_1 \cap i_2)jk}}{\sum_{k=1}^{t_j} p'_{(i_1 \cap i_2)jk}}. \tag{3.18}$$

Note that $\sum_{k=1}^{t_j} p^*_{(i_1 \cup i_2)jk} = \sum_{k=1}^{t_j} p^*_{(i_1 \cap i_2)jk} = 1$. On the contrary, the standard deviations of both union and intersection use $p'_{(i_1 \cup i_2)jk}$ and $p'_{(i_1 \cap i_2)jk}$, respectively, because they should satisfy $S_{(i_1 \cup i_2)j} \geq \max\{S_{i_1j}, S_{i_2j}\}$ and $S_{(i_1 \cap i_2)j} \leq \min\{S_{i_1j}, S_{i_2j}\}$, respectively, to be used as components of dissimilarity measures.

Definition 3.7 *Let $y'_{(i_1 \cup i_2)j}$ be the union between two transformed histogram-valued observations y'_{i_1j} and y'_{i_2j}. Then, the standard deviation of their union $y'_{(i_1 \cup i_2)j}$ is defined by*

$$S_{(i_1 \cup i_2)j} = \sqrt{\sum_{k=1}^{t_j} \left\{ \frac{(b_{jk} - M^*_{\cup})^2 + (b_{jk} - M^*_{\cup})(b_{j,k+1} - M^*_{\cup}) + (b_{j,k+1} - M^*_{\cup})^2}{3} \right\} p'_{(i_1 \cup i_2)jk}}\ , \tag{3.19}$$

where $M^*_{\cup} = M^*_{(i_1 \cup i_2)j}$; and the standard deviation of the intersection is given by

$$S_{(i_1 \cap i_2)j} = \sqrt{\sum_{k=1}^{t_j} \left\{ \frac{(b_{jk} - M^*_{\cap})^2 + (b_{jk} - M^*_{\cap})(b_{j,k+1} - M^*_{\cap}) + (b_{j,k+1} - M^*_{\cap})^2}{3} \right\} p'_{(i_1 \cap i_2)jk}}\ , \tag{3.20}$$

where $M^*_{\cap} = M^*_{(i_1 \cap i_2)j}$.

Now we can extend the Gowda-Diday (1991a) and Ichino-Yaguchi (1994) dissimilarity measures originally developed for interval-valued data to histogram-valued data.

Definition 3.8 *The* ***extended Gowda-Diday dissimilarity measure*** *between the two transformed histogram-valued observations* $\mathbf{y}'_{i_1}$ *and* $\mathbf{y}'_{i_2}$ *is given by*

$$D_{GD}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \sum_{j=1}^{p} \left[D_{1j}(y'_{i_1j}, y'_{i_2j}) + D_{2j}(y'_{i_1j}, y'_{i_2j}) + D_{3j}(y'_{i_1j}, y'_{i_2j}) \right], \tag{3.21}$$

where for the variable Y_j*, each component of the measure is*

$$\begin{aligned} D_{1j}(y'_{i_1j}, y'_{i_2j}) &= \frac{|S_{i_1j} - S_{i_2j}|}{S_{i_1j} + S_{i_2j}}, \\ D_{2j}(y'_{i_1j}, y'_{i_2j}) &= \frac{S_{i_1j} + S_{i_2j} - 2S_{(i_1 \cap i_2)j}}{S_{i_1j} + S_{i_2j}}, \\ D_{3j}(y'_{i_1j}, y'_{i_2j}) &= \frac{|M_{i_1j} - M_{i_2j}|}{\Psi_j}, \end{aligned}$$

where S_{ij} *is the standard deviation for* y'_{ij} *of Equation (3.14) and* $S_{(i_1 \cap i_2)j}$ *is the standard deviation of the intersection between* y'_{i_1j} *and* y'_{i_2j} *of Equation (3.20),* M_{ij} *is the mean of* y'_{ij} *of Equation (3.13), and where* $\Psi_j = b_{j,t_j+1} - b_{j1}$.

Similarly to the case of interval-valued data, the extension of the Gowda-Diday dissimilarity measure to histogram-valued data is also comprised of three components. The first component, $D_{1j}(y'_{i_1j}, y'_{i_2j})$, relates to the relative size, $D_{2j}(y'_{i_1j}, y'_{i_2j})$ indicates the relative content, and $D_{3j}(y'_{i_1j}, y'_{i_2j})$ measures the relative location. However, unlike interval-valued data, histogram-valued data use the standard deviations (S_{i_1j}, S_{i_2j}, $S_{(i_1 \cap i_2)j}$) to measure relative size and relative content, and the means (M_{i_1j}, M_{i_2j}) as a relative location measure. It is reasonable to use these standard deviations and means because a histogram-valued observation has information for the distribution such as center and dispersion. For each variable, each component has a value between 0 and 1. In addition, as the degree of similarity between two objects increases, the extended Gowda-Diday measure for each variable is closer to 0.

Example 3.2 *Consider the transformed histogram-valued data of Table 3.3. We calculate the extended Gowda-Diday dissimilarity measure for the three species of iris. In order to* *obtain the three components of this measure, we first have to compute the intersection of* $\mathbf{y}'_{i_1} = \mathbf{y}'_1$ *and* $\mathbf{y}'_{i_2} = \mathbf{y}'_2$. *For the variable* $Y_j = Y_1$,

$$for\ i_1 = 1,\ i_2 = 2,\ j = 1,\ k = 1,\ p'_{(i_1 \cap i_2)jk} = p'_{(1\cap 2)11} = \min\{0, 0\} = 0.$$

where subscripts i_1 *and* i_2 *are the* i_1^{th} *and* i_2^{th} *observations, respectively,* j *is the* j^{th} *variable, and* k *is the* k^{th} *subinterval or relative frequency. Similarly,*

$$\begin{aligned} &p'_{(1\cap 2)12} = \min\{0, 0\} = 0,\ p'_{(1\cap 2)13} = \min\{0.36, 0\} = 0, \\ &p'_{(1\cap 2)14} = \min\{0.56, 0.08\} = 0.08,\ p'_{(1\cap 2)15} = \min\{0.08, 0.46\} = 0.08, \\ &p'_{(1\cap 2)16} = \min\{0, 0.40\} = 0,\ p'_{(1\cap 2)17} = \min\{0, 0.06\} = 0. \end{aligned}$$

Thus, for $i_1 = 1,\ i_2 = 2,\ j = 1,\ y'_{(i_1\cap i_2)j} = y'_{(1\cap 2)1}$ *is*

$$\begin{aligned} y'_{(1\cap 2)1} &= \{[0, 0.4), 0;\ [0.4, 0.8), 0;\ [0.8, 1.2), 0;\ [1.2, 1.6), 0.08; \\ &\quad [1.6, 2.0), 0.08;\ [2.0, 2.4), 0;\ [2.4, 2.8), 0\}. \end{aligned}$$

Now we calculate the mean and standard deviation of $y'_{i_1j} = y'_{11}$. *These are, respectively,*

$$M_{i_1j} = M_{11} = \frac{1}{2}\big\{(0 + 0.4)0 + (0.4 + 0.8)0 + (0.8 + 1.2)0.36 + \cdots + (2.4 + 2.8)0\big\} = 1.288,$$

and

$$\begin{aligned} S_{i_1j} = S_{11} &= \left\{0 + 0 + \left(\frac{(-0.488)^2 + (-0.488)(-0.088) + (-0.088)^2}{3}\right) 0.36 + \cdots + 0\right\}^{1/2} \\ &= 0.267. \end{aligned}$$

Similarly, $M_{i_2j} = M_{21} = 1.976$, $S_{i_2j} = S_{21} = 0.312$.

In order to compute $S_{(i_1\cap i_2)j} = S_{(1\cap 2)1}$, *first of all, we obtain* $p^*_{(i_1\cap i_2)jk} = p^*_{(1\cap 2)1k}$, $k = 1, \ldots, 7$; *and then calculate* $M^*_{(i_1\cap i_2)j} = M^*_{(1\cap 2)1}$ *using Equation (3.17). Thus,*

$$\sum_{k=1}^{7} p'_{(1\cap 2)1k} = 0 + 0 + 0 + 0.08 + 0.08 + 0 + 0 = 0.16.$$

Hence, from Equation (3.18), the standardized relative frequencies $p^*_{(i_1 \cap i_2)jk} = p^*_{(1\cap 2)1k}$, $k = 1, \dots, 7$, *are*

$$p^*_{(1\cap 2)11} = \frac{0}{0.16} = 0,\ p^*_{(1\cap 2)12} = 0,\ p^*_{(1\cap 2)13} = 0,\ p^*_{(1\cap 2)14} = \frac{0.08}{0.16} = 0.5,$$
$$p^*_{(1\cap 2)15} = \frac{0.08}{0.16} = 0.5,\ p^*_{(1\cap 2)16} = 0,\ p^*_{(1\cap 2)17} = 0,$$

and

$$M^*_{(i_1\cap i_2)j} = M^*_{(1\cap 2)1} = \frac{1}{2}\big\{0 + \cdots + (1.2 + 1.6)0.5 + (1.6 + 2.0)0.5 + 0\big\} = 1.6.$$

Hence, the standard deviation for the intersection of $y'_{i_1 j} = y'_{11}$ *and* $y'_{i_2 j} = y'_{21}$ *is, from Equation (3.20),*

$$\begin{aligned} S_{(i_1\cap i_2)j} = S_{(1\cap 2)1} &= \Bigg\{0 + \cdots + \left(\frac{(1.2-1.6)^2 + 0 + 0}{3}\right)0.08 \\ &\qquad + \left(\frac{0 + 0 + (2-1.6)^2}{3}\right)0.08 + \cdots + 0\Bigg\}^{1/2} \\ &= 0.092. \end{aligned}$$

By using these mean and standard deviation values, the three components in Equation (3.21) of the extended Gowda-Diday measure between observations $\mathbf{y}'_1$ *and* $\mathbf{y}'_2$ *for variable* Y_1 *can be obtained as follows:*

$$\begin{aligned} D_{11}(y'_{11}, y'_{21}) &= \frac{|0.267 - 0.312|}{0.267 + 0.312} = 0.079, \\ D_{21}(y'_{11}, y'_{21}) &= \frac{0.267 + 0.312 - 2(0.092)}{0.267 + 0.312} = 0.681, \\ D_{31}(y'_{11}, y'_{21}) &= \frac{|1.288 - 1.976|}{2.8 - 0} = 0.246. \end{aligned}$$

Similarly, the three components for variable Y_2 *are* $D_{12}(y'_{12}, y'_{22}) = 0.094$, $D_{22}(y'_{12}, y'_{22}) = 0.794$, *and* $D_{32}(y'_{12}, y'_{22}) = 0.212$. *Thus, the extended Gowda-Diday dissimilarity measure between* $\mathbf{y}'_1$ *and* $\mathbf{y}'_2$ *is given by, from Equation (3.21),*

$$D_{GD}(\mathbf{y}'_1, \mathbf{y}'_2) = (0.079 + 0.681 + 0.246) + (0.094 + 0.794 + 0.212) = 2.106.$$

Similarly, we can calculate these dissimilarities $D_{GD}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2})$ *for all* $i_1, i_2 = 1, 2, 3$*. Hence, we can complete the dissimilarity matrix. The extended Gowda-Diday dissimilarity matrix for the data of Table 3.3 is*

$$\mathbf{D}_{GD} = \begin{pmatrix} 0 & 2.106 & 3.505 \\ 2.106 & 0 & 4.110 \\ 3.505 & 4.110 & 0 \end{pmatrix}.$$

Let us now consider an Ichino-Yaguchi dissimilarity measure for histogram-valued data. For interval-valued data, the Ichino-Yaguchi dissimilarity measure is defined using the Cartesian 'join' and 'meet' functions. In contrast, our extension of the Ichino-Yaguchi measure to histogram-valued data uses the standard deviation of the 'union' and 'intersection' between transformed histogram-valued objects given in Definition 3.7.

Definition 3.9 *The* ***extended Ichino-Yaguchi dissimilarity measure*** *of the two transformed histogram-valued observations* y'_{i_1j} *and* y'_{i_2j} *on the variable* Y_j *is*

$$\phi(y'_{i_1j}, y'_{i_2j}) = S_{(i_1 \cup i_2)j} - S_{(i_1 \cap i_2)j} + \gamma(2S_{(i_1 \cap i_2)j} - S_{i_1j} - S_{i_2j}), \tag{3.22}$$

where $S_{(i_1 \cup i_2)j}$*,* $S_{(i_1 \cap i_2)j}$*, and* S_{ij} *are the standard deviations defined in Equation (3.19), (3.20), and (3.14), respectively, and where* $0 \leq \gamma \leq 0.5$ *is a prespecified constant.*

Since the extended Ichino-Yaguchi dissimilarity measure is not normalized, it has different units of measurement for each variable and may depend on a large scale variable. Thus, we often need a normalized version.

Definition 3.10 *Let* $\phi(y'_{i_1j}, y'_{i_2j})$ *be the extended Ichino-Yaguchi dissimilarity measure for the two transformed histogram-valued observations* y'_{i_1j} *and* y'_{i_2j} *on the variable* Y_j *as given in Equation (3.22). Then, a* ***normalized extended Ichino-Yaguchi dissimilarity measure*** *is given by*

$$\phi^*(y'_{i_1j}, y'_{i_2j}) = \frac{\phi(y'_{i_1j}, y'_{i_2j})}{V_j}, \tag{3.23}$$

where

$$V_j = \sqrt{\frac{5A_{1j} + 2A_{2j} - 6A_{3j}}{24}}, \tag{3.24}$$

where

$$\begin{aligned} A_{1j} &= b_{j1}^2 + b_{j2}^2 + b_{jt_j}^2 + b_{j,t_j+1}^2, \\ A_{2j} &= b_{j1}b_{j2} + b_{jt_j}b_{j,t_j+1}, \\ A_{3j} &= b_{j1}b_{jt_j} + b_{j1}b_{j,t_j+1} + b_{j2}b_{jt_j} + b_{j2}b_{j,t_j+1}. \end{aligned}$$

where b_{jk} *and* $b_{j,k+1}$, $k = 1, \ldots, t_j$, *are lower and upper limits for transformed subintervals, respectively, defined in Definition 3.3.*

In order to normalize the extended Ichino-Yaguchi measure, we use the maximum standard deviation value of the union between two single histogram-valued observations for a variable Y_j. That is, for a variable Y_j, consider the union of two single observations, $y'_{(i_1 \cup i_2)j}$, with transformed subintervals $[b_{jk}, b_{j,k+1})$, $k = 1, \ldots, t_j$. Then, when the transformed relative frequencies corresponding to the first and last subintervals are one (i.e., $p'_{(i_1 \cup i_2)j1} = p'_{(i_1 \cup i_2)jt_j} = 1$) and the others are zero, the standard deviation of the union, $S_{(i_1 \cup i_2)j}$, is maximized and this maximum value is the V_j of Equation (3.24). By dividing the extended Ichino-Yaguchi measure by V_j, the normalized measure can be obtained and it has a value between 0 and 1.

The extended Ichino-Yaguchi dissimilarity measure can be extended to Minkowski distances. The form of the Minkowski distance is the same as in Equation (2.10).

Definition 3.11 *The* ***generalized Minkowski distance*** *between the two transformed histogram-valued objects* $\mathbf{y}'_{i_1}$ *and* $\mathbf{y}'_{i_2}$ *is*

$$D_M^q(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \Big[\sum_{j=1}^{p} \phi(y'_{i_1j}, y'_{i_2j})^q\Big]^{1/q}, \tag{3.25}$$

where $\phi(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2})$ *is the extended Ichino-Yaguchi dissimilarity measure. The* ***city block distance*** *is a Minkowski distance with order* $q = 1$,

$$D_M^1(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \sum_{j=1}^{p} \phi(y'_{i_1j}, y'_{i_2j}). \tag{3.26}$$

Also, when a Minkowski distance has order $q = 2$, it is called the ***Euclidean distance*** *and becomes*

$$D^2_M(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \Big[\sum_{j=1}^{p} \phi(y'_{i_1j}, y'_{i_2j})^2\Big]^{1/2}. \tag{3.27}$$

If Equation (3.23) is applied to the generalized Minkowski distance, it becomes the normalized Minkowski distance which takes account of the scale of measurement on each variable Y_j as follows:

$$D^q_{NM}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \Big[\sum_{j=1}^{p} \phi^*(y'_{i_1j}, y'_{i_2j})^q\Big]^{1/q}, \tag{3.28}$$

where $\phi^*(\cdot, \cdot)$ is given by Equation (3.23). When the relative importance of variables is considered in a Minkowski distance, we can use a weight for each variable. This measure is called the normalized and weighted Minkowski distance, viz.,

$$D^q_{NWM}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \Big[\sum_{j=1}^{p} \omega_j \big\{\phi^*(y'_{i_1j}, y'_{i_2j})\big\}^q\Big]^{1/q}, \tag{3.29}$$

where ω_j is a weight with $\omega_j > 0$ and $\sum_{j=1}^{p} \omega_j = 1$.

Example 3.3 *Consider the transformed histogram-valued data of Table 3.3. We calculate the extended Ichino-Yaguchi dissimilarity measure for the three species of iris. To calculate the extended Ichino-Yaguchi measure, we first obtain the union between two transformed histogram-valued objects $\mathbf{y}'_{i_1} = \mathbf{y}'_1$ and $\mathbf{y}'_{i_2} = \mathbf{y}_2$. For the variable $Y_j = Y_1$,*

$$for\ i_1 = 1,\ i_2 = 2,\ j = 1,\ k = 1,\ p'_{(i_1 \cup i_2)jk} = p'_{(1\cup 2)11} = \max\{0, 0\} = 0,$$

where subscripts i_1 and i_2 are the i^{th}_1 and i^{th}_2 observations, respectively, j is the j^{th} variable, and k is the k^{th} subinterval or relative frequency. Similarly,

$$\begin{aligned}
&p'_{(1\cup 2)12} = \max\{0, 0\} = 0,\ p'_{(1\cup 2)13} = \max\{0.36, 0\} = 0.36,\\
&p'_{(1\cup 2)14} = \max\{0.56, 0.08\} = 0.56,\ p'_{(1\cup 2)15} = \max\{0.08, 0.46\} = 0.46,\\
&p'_{(1\cup 2)16} = \max\{0, 0.40\} = 0.40,\ p'_{(1\cup 2)17} = \max\{0, 0.06\} = 0.06.
\end{aligned}$$

Thus, the union $y'_{(i_1 \cup i_2)j} = y'_{(1\cup 2)1}$ *is*

$$\begin{aligned} y'_{(1\cup 2)1} &= \big\{[0,0.4),0;\ [0.4,0.8),0;\ [0.8,1.2),0.36;\ [1.2,1.6),0.56; \\ &\qquad [1.6,2.0),0.46;\ [2.0,2.4),0.40;\ [2.4,2.8),0.06\big\}. \end{aligned}$$

In order to compute the standard deviation of union $S_{(i_1\cup i_2)j} = S_{(1\cup 2)1}$*, first of all, we obtain* $p^*_{(1_1\cup i_2)jk} = p^*_{(1\cup 2)1k}$*,* $k = 1,\ldots,7$*; and then calculate the mean of union* $M^*_{(i_1\cup i_2)j} = M^*_{(1\cup 2)1}$ *using Equation (3.15). Hence,*

$$\sum_{k=1}^{7} p'_{(1\cup 2)1k} = 0 + 0 + 0.36 + 0.56 + 0.46 + 0.40 + 0.06 = 1.84.$$

Therefore, from Equation (3.16), the standardized relative frequencies $p^*_{(i_1\cup i_2)jk} = p^*_{(1\cup 2)1k}$*,* $k = 1,\ldots,7$*, are*

$$\begin{aligned} &p^*_{(1\cup 2)11} = 0,\ p^*_{(1\cup 2)12} = 0,\ p^*_{(1\cup 2)13} = \frac{0.36}{1.84} = 0.196, \\ &p^*_{(1\cup 2)14} = \frac{0.56}{1.84} = 0.304,\ p^*_{(1\cup 2)15} = \frac{0.46}{1.84} = 0.250, \\ &p^*_{(1\cup 2)16} = \frac{0.40}{1.84} = 0.217,\ p^*_{(1\cup 2)17} = \frac{0.06}{1.84} = 0.033, \end{aligned}$$

and the mean of union is , from Equation (3.15),

$$M^*_{(i_1\cup i_2)j} = M^*_{(1\cup 2)1} = \frac{1}{2}\big[0 + 0 + (0.8+1.2)0.196 + \cdots + (2.4+2.8)0.033\big] = 1.635.$$

Thus, the standard deviation for the union of $y'_{i_1 j} = y'_{11}$ *and* $y'_{i_2 j} = y'_{21}$ *is, from Equation (3.19),*

$$\begin{aligned} S_{(i_1\cup i_2)j} = S_{(1\cup 2)1} &= \Bigg\{\left(\frac{(-0.835)^2 + (-0.835)(-0.435) + (-0.435)^2}{3}\right)0.36 \\ &\qquad + \cdots + \left(\frac{2.834}{3}\right)0.06\Bigg\}^{1/2} = 0.630. \end{aligned}$$

From Example 3.2, we know $S_{i_1 j} = S_{11}$*,* $S_{i_2 j} = S_{21}$*, and* $S_{(i_1\cap i_2)j} = S_{(1\cap 2)1}$*. Thus, if we assume* $\gamma = 0.5$*, then the extended Ichino-Yaguchi measure* $\phi(y'_{11}, y'_{21})$ *is, from Equation (3.22),*

$$\phi(y'_{11}, y'_{21}) = 0.630 - 0.092 + 0.5(2 \times 0.092 - 0.267 - 0.312) = 0.340.$$

Similarly, $\phi(y'_{12}, y'_{22}) = 0.686.$

Now we compute the normalized extended Ichino-Yaguchi measure for $\phi(y'_{11}, y'_{21})$*. First, from Equation (3.24),*

$$\begin{aligned} A_{1j} = A_{11} &= 0^2 + 0.4^2 + 2.4^2 + 2.8^2 = 13.76, \\ A_{2j} = A_{21} &= 0 \times 0.4 + 2.4 \times 2.8 = 6.72, \\ A_{3j} = A_{31} &= 0 \times 2.4 + 0 \times 2.8 + 0.4 \times 2.4 + 0.4 \times 2.8 = 2.08. \end{aligned}$$

Thus,

$$V_j = V_1 = \sqrt{\frac{5 \times 13.76 + 2 \times 6.72 - 6 \times 2.08}{24}} = 1.705.$$

Therefore, substituting into Equation (3.23), we obtain

$$\phi^*(y'_{11}, y'_{21}) = \frac{0.340}{1.705} = 0.200.$$

Similarly, $\phi^*(y'_{12}, y'_{22}) = 0.162.$

By using $\phi^*(y'_{11}, y'_{21}) = 0.200$ *and* $\phi^*(y'_{12}, y'_{22}) = 0.162$*, the normalized Euclidean distance of* $\mathbf{y}'_1$ *and* $\mathbf{y}'_2$ *can be computed, from Equation (3.28) with* $q = 2$*, as follows:*

$$D^2_{NM}(\mathbf{y}_1, \mathbf{y}_2) = \left[0.200^2 + 0.162^2\right]^{1/2} = 0.257.$$

Similarly, the normalized Euclidean distances for all $i_1, i_2 = 1, 2, 3$ *can be calculated. Hence, the normalized Euclidean distance matrix can be completed to give*

$$\mathbf{D}^2_{NM} = \begin{pmatrix} 0 & 0.257 & 0.533 \\ 0.257 & 0 & 0.871 \\ 0.533 & 0.871 & 0 \end{pmatrix}.$$

As indicated, the extended Gowda-Diday and extended Ichino-Yaguchi measures for histogram-valued data are extensions of their counterparts for interval-valued data. We now introduce a new measure based on histogram-valued data, as follows. Since histogram-valued data can be considered as probability density functions, cumulative density functions can be obtained from the data. Thus, the area between two cumulative density functions can be used as a dissimilarity measure.

Definition 3.12 *The* ***Cumulative Density Function (CDF) dissimilarity measure*** *between the two histogram-valued objects,* $\mathbf{y}_{i_1}$ *and* $\mathbf{y}_{i_2}$*, is defined by*

$$D_{CDF}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \sum_{j=1}^{p} D_{CDF}(y'_{i_1j}, y'_{i_2j}) = \sum_{j=1}^{p} \left[\sum_{k=1}^{t_j} \left\{ T_j \left| F_{i_1jk} - F_{i_2jk} \right| \right\} \right], \tag{3.30}$$

where T_j *is the length of a transformed subinterval (i.e.,* $T_j = b_{j,k+1} - b_{jk}$*), and* $F_{ijt} = \sum_{k=1}^{t} p'_{ijk}$ *which is a cumulative relative frequency. The* ***normalized CDF dissimilarity measure*** *which does not depend on units of measurement for variables is given by*

$$D_{NCDF}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2}) = \sum_{j=1}^{p} D_{NCDF}(y'_{i_1j}, y'_{i_2j}) = \sum_{j=1}^{p} \left[\Psi_j^{-1} \sum_{k=1}^{t_j} \left\{ T_j \left| F_{i_1jk} - F_{i_2jk} \right| \right\} \right], \tag{3.31}$$

where $\Psi_j = b_{j,t_j+1} - b_{j1}$.

The dissimilarity measure $D_{NCDF}(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ has a value between 0 and p, where p is the number of variables.

Example 3.4 *From the transformed histogram-valued data of Table 3.3, we calculate the CDF dissimilarity measure for the three species of iris. First of all, the cumulative relative frequencies for* $y'_{i_1j} = y'_{11}$ *are obtained by*

$$for\ i_1 = 1,\ j = 1,\ k = 1,\ F_{i_1jk} = F_{111} = p'_{111} = 0,$$

where subscripts i_1 *is the* i_1^{th} *observation,* j *is the* j^{th} *variable, and* k *is the* k^{th} *subinterval or relative frequency. Similarly,*

$$F_{112} = F_{111} + p'_{112} = 0 + 0 = 0,$$

$$F_{113} = F_{112} + p'_{113} = 0 + 0.36 = 0.36,$$

$$F_{114} = F_{113} + p'_{114} = 0.36 + 0.56 = 0.92,$$

$$F_{115} = F_{114} + p'_{115} = 0.92 + 0.08 = 1,$$

$$F_{116} = F_{115} + p'_{116} = 1 + 0 = 1,$$

$$F_{117} = F_{116} + p'_{117} = 1 + 0 = 1.$$

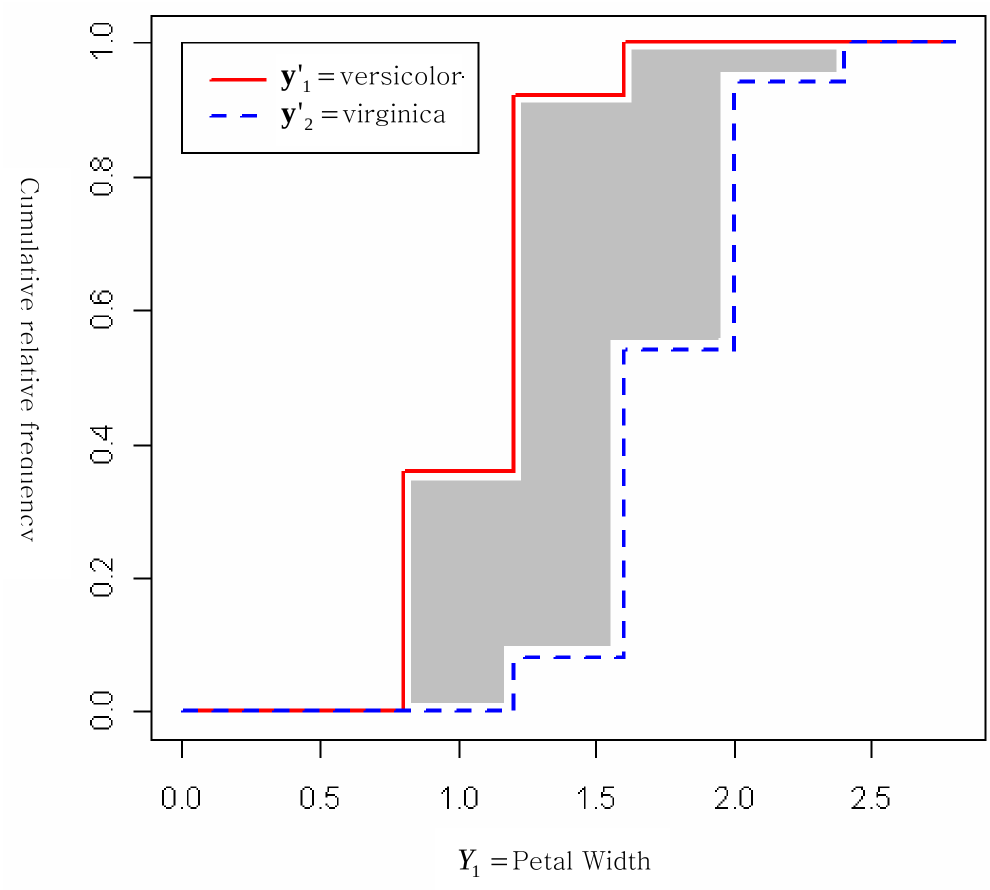


Figure 3.4: Plot of the CDF dissimilarity value between y'_{11} and y'_{21}.

Similarly, cumulative relative frequencies for $y'_{i_2j} = y'_{21}$ *are*

$$F_{211} = 0,\ \ F_{212} = 0,\ \ F_{213} = 0,\ \ F_{214} = 0.08,\ \ F_{215} = 0.54,\ \ F_{216} = 0.94,\ \ F_{217} = 1.$$

Thus, since $T_j = T_1 = b_{1,k+1} - b_{1k} = 0.4$ *for* $k = 1, \ldots, t_1$*, the CDF dissimilarity measure of* y'_{11} *and* y'_{21} *is obtained by*

$$D_{CDF}(y'_{11}, y'_{21}) = 0.4\Big[0 + 0 + \big|0.36 - 0\big| + \big|0.92 - 0.08\big| + \cdots + \big|1 - 1\big|\Big] = 0.688.$$

The shaded area of Figure 3.4 is the CDF dissimilarity measure of $y'_{i_1j} = y'_{11}$ *and* $y_{i_2j} = y'_{21}$*. Similarly,* $D_{CDF}(y'_{12}, y'_{22}) = 1.38$*. Thus, the CDF dissimilarity measure of* $\mathbf{y}_{i_1} = \mathbf{y}_1$ *and*

$\mathbf{y}_{i_2} = \mathbf{y}_2$ *is, from Equation (3.30),*

$$D_{CDF}(\mathbf{y}'_1, \mathbf{y}'_2) = 0.688 + 1.38 = 2.068.$$

In addition, from Equation (3.31), the normalized CDF dissimilarity measure of $\mathbf{y}'_1$ *and* $\mathbf{y}'_2$ *is calculated as follows:*

$$D_{NCDF}(\mathbf{y}'_1, \mathbf{y}'_2) = \left(\frac{0.688}{2.8}\right) + \left(\frac{1.38}{6.5}\right) = 0.458.$$

Similarly, we can calculate these distances $D_{NCDF}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2})$ *for all* $i_1, i_2 = 1, 2, 3$*. Hence, we can complete the normalized CDF dissimilarity measure for the histogram-valued data of Table 3.2 as follows:*

$$\mathbf{D}_{NCDF} = \begin{pmatrix} 0 & 0.458 & 0.817 \\ 0.458 & 0 & 1.275 \\ 0.817 & 1.275 & 0 \end{pmatrix}.$$

3.4 Hierarchical Divisive Clustering Methods

In this section, we show how to extend the Chavent (1998, 2000) monothetic method reviewed in Section 2.4 to histogram-valued data (see Section 3.4.1), and introduce a polythetic method for symbolic objects (in Section 3.4.2). Both the monothetic and polythetic algorithms are hierarchical divisive clustering methods. The difference between the two algorithms is in the strategy used to find the bipartition minimizing the within-cluster variance. The former uses a single variable at each stage. In contrast, the latter uses all variables simultaneously at each stage.

Suppose that there are p-dimensional histogram-valued random variables $\{Y_j,\ j = 1, \ldots, p\}$ with observations $\mathbf{y}_i \in \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i = \big\{y_{ij},\ j = 1, \ldots, p\big\}$ and P_r is a partition of Ω at the r^{th} stage. Then, $P_r = \{C_u,\ u = 1, \ldots, r\}$, where $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$ is a cluster of size n_u, where $\sum_{u=1}^{r} n_u = n$. At the $(r+1)^{th}$ stage, a single cluster C_u in P_r is bisected into C_u^1 and C_u^2. Thus, a new partition can be written as

$$P_{r+1} = \big(P_r \cup \{C_u^1, C_u^2\}\big) - \{C_u\}. \tag{3.32}$$

Now our interest is which cluster C_u is bisected and how to find the optimal C_u^1 and C_u^2 for that C_u. If we have a partition P_r, there are $\sum_{u=1}^{r}(2^{n_u-1}-1)=z$ (say) possible $(r+1)^{th}$ bipartitions of P_{r+1}. Since the number of possible bipartitions exponentially increases as the number of objects increases, we may not be able to examine all possible bipartitions due to computational time and cost. In order to solve this problem, we need a criterion and strategy which can find the optimal C_u^1 and C_u^2 without having to consider all z possibilities. The criterion used in this study is minimizing the within-cluster variance and maximizing the between-cluster variance.

Definition 3.13 *For a cluster $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$, the **within-cluster variance** $I(C_u)$ is, for $u = 1, \ldots, r$,*

$$I(C_u) = \frac{1}{2\tau}\sum_{i_1=1}^{n_u}\sum_{i_2=1}^{n_u} w_{i_1}w_{i_2}D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}),\ i_1, i_2 = 1, \ldots, n_u, \tag{3.33}$$

where $D(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ is a dissimilarity or distance measure between the observation $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ in C_u and where w_i is the weight for the object $\mathbf{y}_i$, and where $\tau = \sum_{i=1}^{n_u} w_i$.

If the weight is the same for each observation, the weight w_i is equal to 1 or equal to $1/n$, where n is the total number of objects in Ω. If $w_i = 1$ for all i, the within-cluster variance can be written as

$$I(C_u) = \frac{1}{2n}\sum_{i_1=1}^{n_u}\sum_{i_2=1}^{n_u} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}). \tag{3.34}$$

If $w_i = 1/n$, the $I(C_u)$ is

$$I(C_u) = \frac{1}{2nn_u}\sum_{i_1=1}^{n_u}\sum_{i_2=1}^{n_u} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}). \tag{3.35}$$

Definition 3.14 *For a partition $P_r = (C_1, \ldots, C_r)$ at the r^{th} stage, the **total within-cluster variance** is given by, for $r = 1, \ldots, n-1$,*

$$W(P_r) = \sum_{u=1}^{r} I(C_u). \tag{3.36}$$

Definition 3.15 *For a partition $P_r = (C_1, \ldots, C_r)$ at the r^{th} stage, where $\Omega \equiv P_1$, the* ***between-cluster variance*** *is defined by, for $r = 1, \ldots, n-1$,*

$$B(P_r) = W(\Omega) - W(P_r). \tag{3.37}$$

From Equation (3.32), the total within-cluster variance for P_{r+1} can be written as

$$W(P_{r+1}) = W(P_r) - \{I(C_u) - I(C_u^1) - I(C_u^2)\}. \tag{3.38}$$

Thus, minimizing $W(P_{r+1})$ is equivalent to maximizing the decrement value of the within-cluster variance Δ_u, where

$$\Delta_u = \big\{I(C_u) - I(C_u^1) - I(C_u^2)\big\}. \tag{3.39}$$

Therefore, when we want to minimize $W(P_{r+1})$, we have to find the bipartition (C_u^1, C_u^2) that maximizes Δ_u.

3.4.1 Monothetic Algorithm

A monothetic algorithm uses a single variable to find the bipartition that minimizes the within-cluster variance. It can also find a binary question for a single variable which shows a monothetic characteristic at each stage. The form of a binary question is 'Is $Y_j \leq c_r$?', where c_r is the cut point at the r^{th} stage.

Suppose that there are p-dimensional histogram-valued random variables $\{Y_j,\ j = 1, \ldots, p\}$ with observations $\mathbf{y}_i \in \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ with $\mathbf{y}_i = \big\{y_{ij},\ j = 1, \ldots, p\big\} = \big\{[a_{ijk}, a_{ij,k+1}),\ p_{ijk},\ j = 1, \ldots, p,\ k = 1, \ldots, v_{ij}\big\}$. Then, we can obtain transformed histogram-valued observations $\mathbf{y}'_i = \big\{y'_{ij},\ j = 1, \ldots, p\big\} = \big\{[b_{ijk}, b_{ij,k+1}),\ p'_{ijk},\ j = 1, \ldots, p,\ k = 1, \ldots, t_j\big\}$, $i = 1, \ldots, n$, from the original histogram-valued observations $\mathbf{y}_i,\ i = 1, \ldots, n$, using Definition 3.3; and from these transformed histogram-valued observations, we can calculate distance or dissimilarity values among all observations such as, the extended Gowda-Diday, the extended Ichino-Yaguchi, and the CDF dissimilarity measures proposed in Section 3.3. Using these transformed observations and dissimilarity values, the monothetic clustering method can be achieved.

The monothetic algorithm for histogram-valued data is very similar to the case of interval-valued data introduced in Section 2.4. By adapting that algorithm (developed by Chavent, 1998), we obtain an extension of the algorithm for histogram-valued data as follows:

- **Step 1**: Start with $P_1 \equiv \Omega = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$. Then, $r = 1$.

- **Step 2**: At the r^{th} stage, we have a partition $P_r = \{C_u,\ u = 1, \dots, r\}$, and each cluster contains histogram-valued objects (i.e., $C_u = \{\mathbf{y}_i,\ i = 1, \dots, n_u\}$).

- **Step 3**: For a variable Y_j, calculate the mean M_{ij} for the transformed observations y'_{ij} of all objects in each cluster using Equation (3.13); and then sort the objects in C_u in ascending order using the value of M_{ij}. Let $\{\mathbf{y}^j_{(i)},\ i = 1, \dots, n_u\}$ be the sorted objects in C_u for the variable Y_j, and let $C^1_u = \{\mathbf{y}^j_{(1)}, \dots, \mathbf{y}^j_{(l)}\}$ and $C^2_u = \{\mathbf{y}^j_{(l+1)}, \dots, \mathbf{y}^j_{(n_u)}\}$, $l = 1, \dots, n_u - 1$. Then, there are $\sum_{u=1}^{r}(n_u - 1)$ possible bipartitions in the partition P_r. For the variable Y_j, find the C_u and its bipartition (C^1_u, C^2_u) with the smallest total within-cluster variance of the $\sum_{u=1}^{r}(n_u - 1)$ possible bipartitions by maximizing Δ_u defined in Equation (3.39); and then set $C^j_u = C_u$, $C^{1j}_u = C^1_u$ and $C^{2j}_u = C^2_u$.

- **Step 4**: For all the variables $Y_j, j = 1, \dots, p$, repeat Step 3. Then, we can find the optimal (C^1_u, C^2_u) and j satisfying
$$\max_j \{I(C^j_u) - I(C^{1j}_u) - I(C^{2j}_u)\}.$$

- **Step 5**: In order to obtain the cut point c_r, calculate the mean of the union between $y'_{(l)j}$ and $y'_{(l+1)j}$ for the identified (C^1_u, C^2_u) and j in Step 4 using Equation (3.15). That is, the cut point c_r is
$$c_r = M^*_{(l \cup l+1)j}. \tag{3.40}$$
Thus, the binary question at the r^{th} stage is 'Is $Y_j \leq c_r$?'.

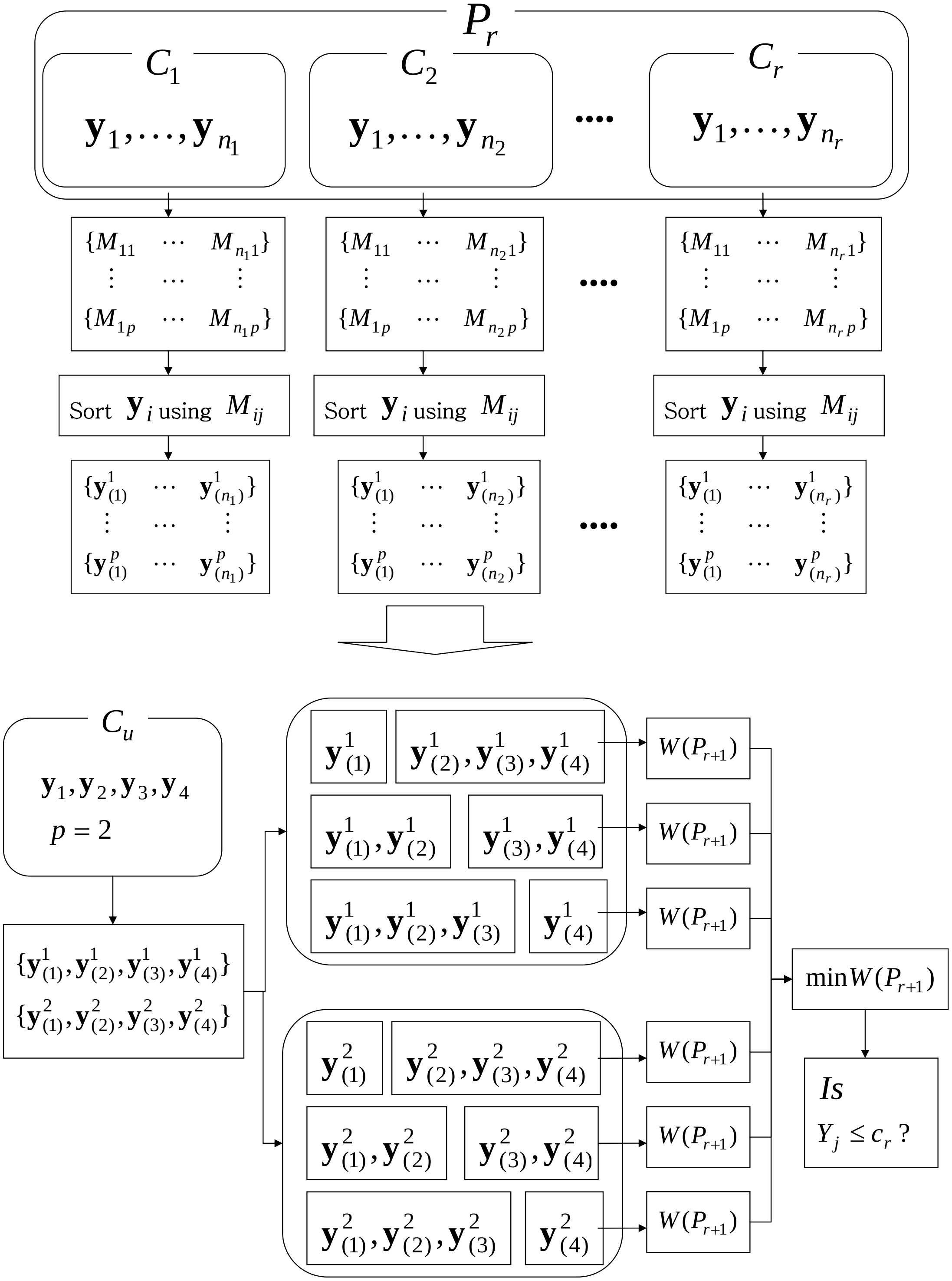

P_r
C_1
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_1}$
C_2
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_2}$
....
C_r
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_r}$
$\{M_{11} \cdots M_{n_1 1}\}$ $\vdots \cdots \vdots$ $\{M_{1p} \cdots M_{n_1 p}\}$
$\{M_{11} \cdots M_{n_2 1}\}$ $\vdots \cdots \vdots$ $\{M_{1p} \cdots M_{n_2 p}\}$
$\{M_{11} \cdots M_{n_r 1}\}$ $\vdots \cdots \vdots$ $\{M_{1p} \cdots M_{n_r p}\}$
Sort $\mathbf{y}_i$ using M_{ij}
Sort $\mathbf{y}_i$ using M_{ij}
Sort $\mathbf{y}_i$ using M_{ij}
$\{\mathbf{y}^1_{(1)} \cdots \mathbf{y}^1_{(n_1)}\}$ $\vdots \cdots \vdots$ $\{\mathbf{y}^p_{(1)} \cdots \mathbf{y}^p_{(n_1)}\}$
$\{\mathbf{y}^1_{(1)} \cdots \mathbf{y}^1_{(n_2)}\}$ $\vdots \cdots \vdots$ $\{\mathbf{y}^p_{(1)} \cdots \mathbf{y}^p_{(n_2)}\}$
$\{\mathbf{y}^1_{(1)} \cdots \mathbf{y}^1_{(n_r)}\}$ $\vdots \cdots \vdots$ $\{\mathbf{y}^p_{(1)} \cdots \mathbf{y}^p_{(n_r)}\}$
C_u
$\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\mathbf{y}_4$
$p=2$
$\{\mathbf{y}^1_{(1)},\mathbf{y}^1_{(2)},\mathbf{y}^1_{(3)},\mathbf{y}^1_{(4)}\}$
$\{\mathbf{y}^2_{(1)},\mathbf{y}^2_{(2)},\mathbf{y}^2_{(3)},\mathbf{y}^2_{(4)}\}$
$\mathbf{y}^1_{(1)}$ | $\mathbf{y}^1_{(2)},\mathbf{y}^1_{(3)},\mathbf{y}^1_{(4)}$
$\mathbf{y}^1_{(1)},\mathbf{y}^1_{(2)}$ | $\mathbf{y}^1_{(3)},\mathbf{y}^1_{(4)}$
$\mathbf{y}^1_{(1)},\mathbf{y}^1_{(2)},\mathbf{y}^1_{(3)}$ | $\mathbf{y}^1_{(4)}$
$\mathbf{y}^2_{(1)}$ | $\mathbf{y}^2_{(2)},\mathbf{y}^2_{(3)},\mathbf{y}^2_{(4)}$
$\mathbf{y}^2_{(1)},\mathbf{y}^2_{(2)}$ | $\mathbf{y}^2_{(3)},\mathbf{y}^2_{(4)}$
$\mathbf{y}^2_{(1)},\mathbf{y}^2_{(2)},\mathbf{y}^2_{(3)}$ | $\mathbf{y}^2_{(4)}$
$W(P_{r+1})$
$W(P_{r+1})$
$W(P_{r+1})$
$W(P_{r+1})$
$W(P_{r+1})$
$W(P_{r+1})$
$\min W(P_{r+1})$
Is
$Y_j \le c_r$?

Figure 3.5: The flow chart for the monothetic algorithm.

Table 3.4: Ruspini histogram-valued data.

	Y_1	Y_2
$\mathbf{y}_1$	$\{[0,5),0.100;\ [5,10),0.100;\ [10,15),0.200;$ $[15,20),0.100;\ [20,25),0.100;\ [25,30),0.250;$ $[30,35),0.100;\ [35,40),0.050\}$	$\{[45,50),0.050;\ [50,55),0.150;\ [55,60),0.200;$ $[60,65),0.200;\ [65,70),0.050;\ [70,75),0.200;$ $[75,80),0.100;\ [80,85),0.000;\ [85,90),0.050\}$
$\mathbf{y}_2$	$\{[25,30),0.044;\ [30,35),0.217;\ [35,40),0.130;$ $[40,45),0.174;\ [45,50),0.174;\ [50,55),0.174;$ $[55,60),0.044;\ [60,65),0.044\}$	$\{[120,125),0.044;\ [125,130),0.000;\ [130,135),0.000;$ $[135,140),0.087;\ [140,145),0.348;\ [145,150),0.217;$ $[150,155),0.261;\ [155,160),0.044\}$
$\mathbf{y}_3$	$\{[70,80),0.118;\ [80,90),0.177;\ [90,100),0.294;$ $[100,110),0.235;\ [110,120),0.177\}$	$\{[90,95),0.059;\ [95,100),0.118;\ [100,105),0.000;$ $[105,110),0.000;\ [110,115),0.294;\ [115,120),0.235;$ $[120,125),0.118;\ [125,130),0.118;\ [130,135),0.059;$
$\mathbf{y}_4$	$\{[55,60),0.067;\ [60,65),0.267;\ [65,70),0.333;$ $[70,75),0.067;\ [75,80),0.200;\ [80,85),0.067\}$	$\{[0,5),0.067;\ [5,10),0.000;\ [10,15),0.267;$ $[15,20),0.200;\ [20,25),0.267;\ [25,30),0.133;$ $[30,35),0.067\}$

- **Step 6**: Repeat Steps 2–5 until $r = R$ or $r = n$, where R is a prespecified value and n is the number of objects.

Figure 3.5 shows the process of the monothetic algorithm.

Example 3.5 *Consider the Ruspini (1970) histogram-valued data shown in Table 3.4. This classical Ruspini dataset is artificial data and has 75 observations with two variables and it is well known that there are four clusters in this dataset. In this example, we use these four clusters as histogram-valued objects to illustrate the monothetic method. First of all, we transform the histogram-valued data of Table 3.4 using Definition 3.3. These transformed histogram-valued data are shown in Table 3.5. Now we have four transformed histogram-valued objects (i.e.,* $\{\mathbf{y}'_1, \mathbf{y}'_2, \mathbf{y}'_3, \mathbf{y}'_4\}$*) with two variables* Y_1 *and* Y_2*.*

From these transformed data, we can calculate dissimilarity values between two objects. In this example, we use the CDF dissimilarity measure defined in Equation (3.30). Since Y_1 *and* Y_2 *have a similar range of values, we do not use the normalized measure. To obtain the CDF dissimilarity measure for four objects, first of all, the cumulative relative frequencies*

Table 3.5: Ruspini transformed histogram-valued data.

Y_1	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	Y_1	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$
$[b_{1k}, b_{1,k+1})$	p'_{11k}	p'_{21k}	p'_{31k}	p'_{41k}	$[b_{1k}, b_{1,k+1})$	p'_{11k}	p'_{21k}	p'_{31k}	p'_{41k}
$[0, 5)$	0.100	0.000	0.000	0.000	$[5, 10)$	0.100	0.000	0.000	0.000
$[10, 15)$	0.200	0.000	0.000	0.000	$[15, 20)$	0.100	0.000	0.000	0.000
$[20, 25)$	0.100	0.000	0.000	0.000	$[25, 30)$	0.250	0.043	0.000	0.000
$[30, 35)$	0.100	0.217	0.000	0.000	$[35, 40)$	0.050	0.130	0.000	0.000
$[40, 45)$	0.000	0.174	0.000	0.000	$[45, 50)$	0.000	0.174	0.000	0.000
$[50, 55)$	0.000	0.174	0.000	0.000	$[55, 60)$	0.000	0.043	0.000	0.067
$[60, 65)$	0.000	0.043	0.000	0.267	$[65, 70)$	0.000	0.000	0.000	0.333
$[70, 75)$	0.000	0.000	0.059	0.067	$[75, 80)$	0.000	0.000	0.059	0.200
$[80, 85)$	0.000	0.000	0.088	0.067	$[85, 90)$	0.000	0.000	0.088	0.000
$[90, 95)$	0.000	0.000	0.147	0.000	$[95, 100)$	0.000	0.000	0.147	0.000
$[100, 105)$	0.000	0.000	0.118	0.000	$[105, 110)$	0.000	0.000	0.118	0.000
$[110, 115)$	0.000	0.000	0.088	0.000	$[115, 120)$	0.000	0.000	0.088	0.000
Y_2	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	Y_2	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$
$[b_{2k}, b_{2,k+1})$	p'_{12k}	p'_{22k}	p'_{32k}	p'_{42k}	$[b_{2k}, b_{2,k+1})$	p'_{12k}	p'_{22k}	p'_{32k}	p'_{42k}
$[0, 5)$	0.000	0.000	0.000	0.067	$[5, 10)$	0.000	0.000	0.000	0.000
$[10, 15)$	0.000	0.000	0.000	0.267	$[15, 20)$	0.000	0.000	0.000	0.200
$[20, 25)$	0.000	0.000	0.000	0.267	$[25, 30)$	0.000	0.000	0.000	0.133
$[30, 35)$	0.000	0.000	0.000	0.067	$[35, 40)$	0.000	0.000	0.000	0.000
$[40, 45)$	0.000	0.000	0.000	0.000	$[45, 50)$	0.050	0.000	0.000	0.000
$[50, 55)$	0.150	0.000	0.000	0.000	$[55, 60)$	0.200	0.000	0.000	0.000
$[60, 65)$	0.200	0.000	0.000	0.000	$[65, 70)$	0.050	0.000	0.000	0.000
$[70, 75)$	0.200	0.000	0.000	0.000	$[75, 80)$	0.100	0.000	0.000	0.000
$[80, 85)$	0.000	0.000	0.000	0.000	$[85, 90)$	0.050	0.000	0.000	0.000
$[90, 95)$	0.000	0.000	0.059	0.000	$[95, 100)$	0.000	0.000	0.118	0.000
$[100, 105)$	0.000	0.000	0.000	0.000	$[105, 110)$	0.000	0.000	0.000	0.000
$[110, 115)$	0.000	0.000	0.294	0.000	$[115, 120)$	0.000	0.000	0.235	0.000
$[120, 125)$	0.000	0.043	0.118	0.000	$[125, 130)$	0.000	0.000	0.118	0.000
$[130, 135)$	0.000	0.000	0.059	0.000	$[135, 140)$	0.000	0.087	0.000	0.000
$[140, 145)$	0.000	0.348	0.000	0.000	$[145, 150)$	0.000	0.217	0.000	0.000
$[150, 155)$	0.000	0.261	0.000	0.000	$[155, 160)$	0.000	0.043	0.000	0.000

for $y'_{i_1j} = y'_{11}$, $F_{i_1jk} = F_{11k}$, *are calculated by*

$$\begin{aligned}
&F_{111} = 0.1,\ F_{112} = 0.1 + 0.1 = 0.2,\ F_{113} = 0.2 + 0.2 = 0.4,\\
&F_{114} = 0.4 + 0.1 = 0.5,\ F_{115} = 0.5 + 0.1 = 0.6,\ F_{116} = 0.6 + 0.25 = 0.85,\\
&F_{117} = 0.85 + 0.1 = 0.95,\ F_{118} = 0.95 + 0.05 = 1, \ldots,\ F_{11,24} = 1.
\end{aligned}$$

Similarly, the cumulative relative frequencies for $y'_{i_2j} = y'_{21}$, $F_{i_2jk} = F_{21k}$, *are*

$$\begin{aligned}
&F_{211} = 0,\ F_{212} = 0,\ F_{213} = 0,\ F_{214} = 0,\ F_{215} = 0,\\
&F_{216} = 0.043, F_{217} = 0.261,\ F_{218} = 0.391, \ldots,\ F_{21,24} = 1.
\end{aligned}$$

Thus, since $T_j = T_1 = 5$, *the CDF dissimilarity measure of* $y'_{i_1j} = y'_{11}$ *and* $y'_{i_2j} = y'_{12}$ *is*

$$\begin{aligned}
D_{CDF}(y'_{11}, y'_{21}) &= 5 \times \Big\{ \big|0.1 - 0\big| + \big|0.2 - 0\big| + \big|0.4 - 0\big| + \big|0.5 - 0\big| \\
&\qquad + \big|0.6 - 0\big| + \big|0.85 - 0.043\big| + \big|0.95 - 0.261\big| + \cdots + \big|1 - 1\big| \Big\} \\
&= 23.652.
\end{aligned}$$

Similarly, $D_{CDF}(y'_{12}, y'_{22}) = 81.293$. *Thus, the CDF dissimilarity measure between* $\mathbf{y}'_1$ *and* $\mathbf{y}'_2$ *is*

$$D_{CDF}(\mathbf{y}'_1, \mathbf{y}'_2) = 23.652 + 81.293 = 104.946.$$

Similarly, we can calculate these distances $D_{CDF}(\mathbf{y}'_1, \mathbf{y}'_2)$ *for all* $i_1, i_2 = 1, 2, 3, 4$. *Hence, we can complete the CDF dissimilarity matrix as follows:*

$$\mathbf{D}_{CDF} = \begin{pmatrix} 0 & 104.946 & 127.868 & 94.750 \\ 104.946 & 0 & 84.303 & 152.391 \\ 127.868 & 84.303 & 0 & 123.951 \\ 94.750 & 152.391 & 123.951 & 0 \end{pmatrix}. \tag{3.41}$$

Now, we perform clustering using the monothetic method. At the first stage, we have $P_1 \equiv \Omega = \big\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\big\}$. *To obtain the order of objects, the means of the observations for* *each variable are calculated using Equation (3.13). The mean of* $y'_{ij} = y'_{11}$, M_{11}, *is*

$$M_{ij} = M_{11} = \Big(\frac{0+5}{2}\Big)0.1 + \Big(\frac{5+10}{2}\Big)0.1 + \cdots + \Big(\frac{115+120}{2}\Big)0 = 19.50.$$

Similarly, $M_{ij},\ i = 1,2,3,4,\ j = 1,2$, *can be calculated; Table 3.6 shows the mean values for each variable and each object. From these mean values, the ascending order of objects can be obtained. For variable* Y_1, *we have the order* $\mathbf{y}_1 < \mathbf{y}_2 < \mathbf{y}_4 < \mathbf{y}_3$, *and for variable* Y_2, *the order is* $\mathbf{y}_4 < \mathbf{y}_1 < \mathbf{y}_3 < \mathbf{y}_2$. *Thus, the number of possible bipartitions for variable* Y_1 *is three* $(= n_u - 1 = 4 - 1)$; *these are* $\big(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_4, \mathbf{y}_3\}\big)$, $\big(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_4, \mathbf{y}_3\}\big)$, *and* $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_4\}, \{\mathbf{y}_3\}\big)$. *Similarly, we can obtain the possible bipartitions for variable* Y_2, *viz.,* $\big(\{\mathbf{y}_4\}, \{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_2\}\big)$, $\big(\{\mathbf{y}_4, \mathbf{y}_1\}, \{\mathbf{y}_3, \mathbf{y}_2\}\big)$, *and* $\big(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_3\}, \{\mathbf{y}_2\}\big)$.

Table 3.6: Mean values for each variable and each object.

Variable	Object	$(\mathbf{y}_i)$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$
Y_1	Mean	(M_{i1})	19.50	43.15	96.76	68.83
Y_2	Mean	(M_{i2})	64.25	145.54	114.85	18.83

To find the optimal bipartition, we use these possible bipartitions and the decrement value of the within-cluster variance, Δ_u, *defined in Equation (3.39). The optimal bipartition has the maximum* Δ_u *value. For example, suppose we have a cluster* $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}$ *and a bipartition* $(C_1^1, C_1^2) = (\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\})$. *From the dissimilarity matrix in Equation (3.41), the within-cluster variances* $I(C_1)$, $I(C_1^1)$, *and* $I(C_1^2)$ *are, using Equation (3.35) with weights* $w_i = 1/n$,

$$\begin{aligned} I(C_1) &= \frac{1}{4\times 4}\big\{104.946^2 + 127.868^2 + 94.750^2 + 84.303^2 + 152.391^2 + 123.951^2\big\} \\ &= 5127.203 \\ I(C_1^1) &= \frac{1}{4\times 2}\big\{104.946^2\big\} = 1376.699 \\ I(C_1^2) &= \frac{1}{4\times 2}\big\{123.951^2\big\} = 1920.481. \end{aligned}$$

Thus, the decrement value Δ_1 *is*

$$\Delta_1 = I(C_1) - I(C_1^1) - I(C_1^2) = 5127.203 - 1376.699 - 1920.481 = 1830.023.$$

Similarly, the decrement values for all the possible bipartitions can be calculated as given in Table 3.7.

Table 3.7: Decrement values for the first stage.

Variable	(C_1^1, C_1^2)	$I(C_1^1)$	$I(C_1^2)$	$I(C_1) - I(C_1^1) - I(C_1^2)$
Y_1	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_4, \mathbf{y}_3\})$	0.00	3807.83	1319.37
	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_4, \mathbf{y}_3\})$	1376.70	1920.48	1830.02
	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_4\}, \{\mathbf{y}_3\})$	3601.19	0.00	1526.02
Y_2	$(\{\mathbf{y}_4\}, \{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_2\})$	0.00	2872.56	2254.642
	$(\{\mathbf{y}_4, \mathbf{y}_1\}, \{\mathbf{y}_3, \mathbf{y}_2\})$	1122.20	888.38	**3116.63**
	$(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_3\}, \{\mathbf{y}_2\})$	3390.96	0.00	1736.24

From Table 3.7, we know the bipartition $(\{\mathbf{y}_4, \mathbf{y}_1\}, \{\mathbf{y}_3, \mathbf{y}_2\})$ *has the largest decrement value (3116.63). Note that this bipartition is detected by variable* Y_2*. Thus, the form of the binary question is 'Is* $Y_2 \leq c_1$*?'. From Equation (3.40), the cut point at the first stage,* c_1*, is the mean of the union of the two transformed observations* $\mathbf{y}_1'$ *and* $\mathbf{y}_3'$ *for variable* Y_2*. Thus, from Equation (3.15), the cut point* $c_1 = M^*_{(i_1 \cup i_2)j} = M^*_{(1 \cup 3)2} = 89.55$*, and the binary question at the first stage is 'Is* $Y_2 \leq 89.55$*?'.*

At the second stage, we have the partition $P_2 = \big\{C_1, C_2\big\}$*, where* $C_1 = \{\mathbf{y}_4, \mathbf{y}_1\}$ *and* $C_2 = \{\mathbf{y}_3, \mathbf{y}_2\}$*, and either* C_1 *or* C_2 *is bipartitioned. Similarly to the first stage, we find the optimal bipartition by sorting objects in each cluster by the mean values* M_{ij} *and by using the decrement value of the within-cluster variance defined in Equation (3.39). This result is shown in Table 3.8.*

Table 3.8: Decrement values for the second stage.

Variable	(C_u^1, C_u^2)	$I(C_u^1)$	$I(C_u^2)$	$I(C_u) - I(C_u^1) - I(C_u^2)$
Y_1	$(\{\mathbf{y}_1\}, \{\mathbf{y}_4\})$	0.00	0.00	**1122.20**
	$(\{\mathbf{y}_2\}, \{\mathbf{y}_3\})$	0.00	0.00	888.38
Y_2	$(\{\mathbf{y}_4\}, \{\mathbf{y}_1\})$	0.00	0.00	**1122.20**
	$(\{\mathbf{y}_3\}, \{\mathbf{y}_2\})$	0.00	0.00	888.38

From Table 3.8, we know the optimal bipartition at the second stage is $(\{\mathbf{y}_1\}, \{\mathbf{y}_4\})$*. Unlike the first stage, this optimal bipartition is detected by both variables* Y_1 *and* Y_2*. In this case, to obtain a unique binary question, we choose a variable using dissimilarity values for each*

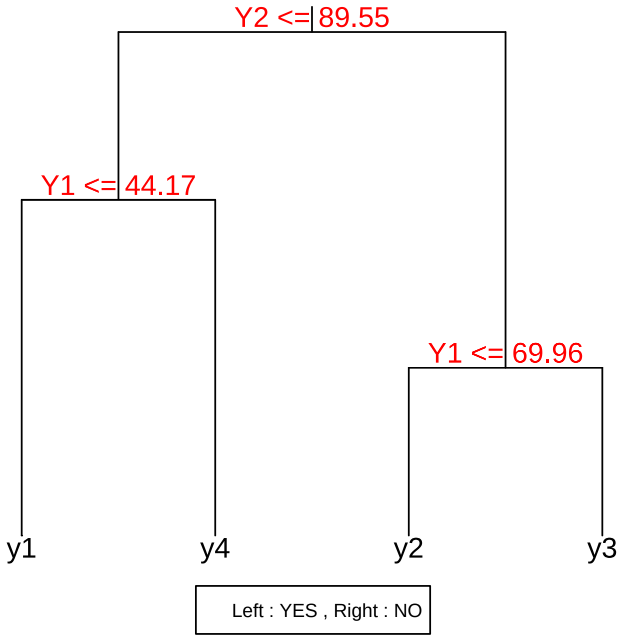


Figure 3.6: Dendrogram for Ruspini data using the monothetic method.

variable. That is, we compare $D_{CDF}(y'_{11}, y'_{41})$ *with* $D_{CDF}(y'_{12}, y'_{42})$*. Since* $D_{CDF}(y'_{11}, y'_{41}) = 49.33 > D_{CDF}(y'_{12}, y'_{42}) = 45.42$*, the form of the binary question at the second stage is 'Is* $Y_1 \leq c_2$*?'. Using Equation (3.15), the cut point is* $c_2 = M^*_{(i_1 \cup i_2)j} = M^*_{(1\cup 4)1} = 44.17$*, and the binary question at the second stage is 'Is* $Y_1 \leq 44.17$*?'.*

Similarly, we can find the optimal bipartition and binary question for the third stage. The dendrogram for the complete clustering result is shown in Figure 3.6. For the first binary question 'Is $Y_2 \leq 89.55$*?', if the answer is 'Yes', then the observations fall into cluster* $\{\mathbf{y}_1, \mathbf{y}_4\}$*, and if 'No' then observations fall into cluster* $\{\mathbf{y}_2, \mathbf{y}_3\}$*.*

As mentioned in Section 3.4, there exist $\sum_{u=1}^{r}(2^{n_u-1} - 1)$ possible bipartitions at the r^{th} stage. However, in the monothetic method, we have $\sum_{u=1}^{r} p(n_u - 1)$ bipartitions because for each cluster, there are $n_u - 1$ bipartitions for each variable due to the fact that objects are sorted in ascending order using the mean; see Step 3. Thus, in order to obtain the optimal bipartition, we need $\sum_{u=1}^{r} p(n_u - 1)$ repetitions of the Step 3 at each stage.

There is a binary question at every stage. This binary question is interpreted to mean that if an object $\mathbf{y}_i$ has 'Yes' as a response for 'Is $Y_j \leq c_r$?', then $\mathbf{y}_i \in C_u^1$, and if 'No', then $\mathbf{y}_i \in C_u^2$. The prespecified number of clusters is R. In Chapter 4, we propose a rule to select the number of clusters. An application of this algorithm is given in Chapter 5; and the program code in software R is provided in the Appendix C.

3.4.2 Polythetic Algorithm

In this section, we propose a polythetic method for clustering symbolic objects including interval-valued, histogram-valued, and multi-valued objects, etc. As explained in Section 3.4.1, the monothetic method depends on a single variable to find the optimal bipartition because it uses objects sorted in ascending mean values for each variable. Thus, this method is simple and has the advantages that it gives both a hierarchy on a dataset and a simple interpretation (i.e., binary questions) for clustering results. However, Chavent (1998) indicated that since it uses a single variable at a time, it might perform poorly in those situations where the cluster structure depends on combinations of variables. In contrast, the polythetic method uses all variables simultaneously through dissimilarity or distance values. That is, this method does not depend on orders of single variables but completely depends on dissimilarity or distance values, and these dissimilarity values consider the dissimilarity for all variables simultaneously.

To avoid considering all possible bipartitions, the method starts by finding the object that is farthest away from the others within clusters $C_u,\ u = 1, \ldots, r$, in a partition P_r. The farthest object would be used as the seed. Let $\mathbf{y}^*$ be a seed and C be the cluster with $\mathbf{y}^*$ that is one of clusters $C_u,\ u = 1, \ldots, r$, in P_r; then C would be bipartitioned into (C^1, C^2). At the beginning, the seed $\mathbf{y}^*$ automatically goes to C^2. That is, $C^1 = C - \{\mathbf{y}^*\}$ and $C^2 = \{\mathbf{y}^*\}$. Then, the cluster C^1 is called the main cluster (or group) and cluster C^2 is the splinter cluster. The method compares whether each object is closer to the main cluster C^1 or to the splinter cluster C^2. If an object is closer to the splinter cluster, it moves into that cluster C^2.

This comparison is iterated until there are no more objects that are closer to the splinter cluster at each stage.

The polythetic algorithm for classical data proposed by MacNaughton-Smith et al. (1964) iteratively uses an average distance between an object and a group of objects. In their method, the object with the maximum average distance between objects in the same cluster is used as a seed and automatically goes into a splinter cluster. Each object in the cluster is compared using average distance between an object in the main cluster and objects in the splinter cluster whether it is close to the main cluster or the splinter cluster, and any objects that are closer to the splinter cluster are moved into the splinter cluster. This step is repeated while, in the main cluster, there are objects that are closer to the splinter cluster.

On the contrary, we propose the polythetic method of symbolic objects using an iterative procedure by a within-cluster variance, instead of an average distance as follows:

- **Step 1**: Let $\mathbf{y}_i$ be a symbolic object. Then, start with $P_1 \equiv \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Then, $r = 1$.

- **Step 2**: At the r^{th} stage, we have a partition $P_r = \{C_u,\ u = 1, \ldots, r\}$, and each cluster has histogram-valued objects (i.e., $C_u = \{\mathbf{y}_i,\ i = 1, \ldots, n_u\}$).

- **Step 3**: Let $\bar{D}_u(\mathbf{y}_i)$ be an average weighted dissimilarity of $\mathbf{y}_i$ for C_u. Then, $\bar{D}_u(\mathbf{y}_i)$ is defined by
$$\bar{D}_u(\mathbf{y}_i) = \frac{\sum_{i' \neq i} \left\{ w_i w_{i'} D(\mathbf{y}_i, \mathbf{y}_{i'}) \right\}}{\tau_u - w_i},\ i = 1, \ldots, n_u, \tag{3.42}$$
where $D(\mathbf{y}_i, \mathbf{y}_{i'})$ is the dissimilarity or distance measure between two symbolic objects $\mathbf{y}_i$ and $\mathbf{y}_{i'}$, $\mathbf{y}_i,\ \mathbf{y}_{i'} \in C_u$ and where w_i is the weight for the $\mathbf{y}_i$ and $\tau_u = \sum_{i=1}^{n_u} w_i$. Let MAD_u be the maximum value of $\bar{D}_u(\mathbf{y}_i)$ for a cluster C_u (i.e., the maximum average dissimilarity). Then, the MAD_u is given by
$$MAD_u = \max_i \{\bar{D}_u(\mathbf{y}_i),\ i = 1, \ldots, n_u\}. \tag{3.43}$$

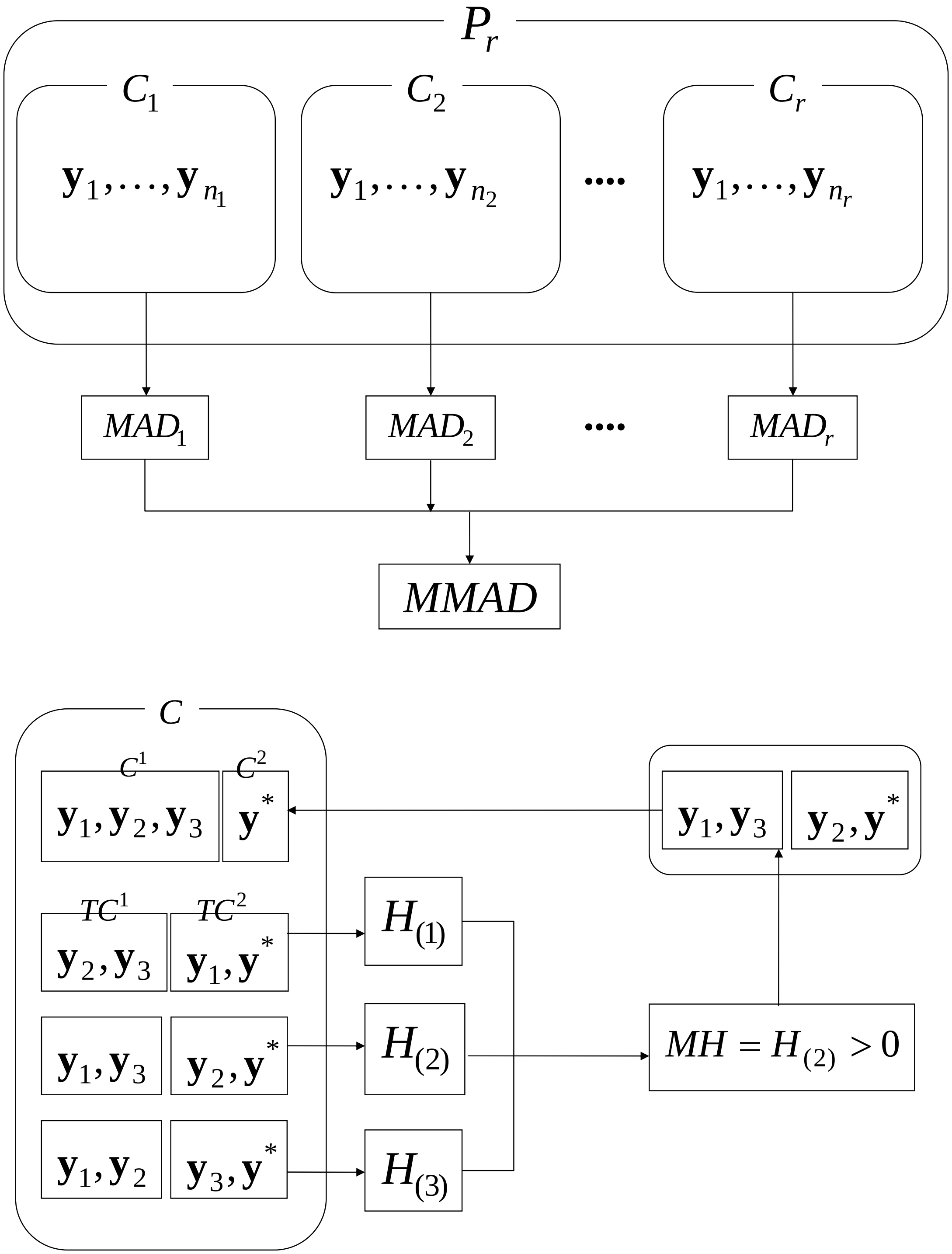

P_r
C_1
C_2
C_r
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_1}$
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_2}$
....
$\mathbf{y}_1,\ldots,\mathbf{y}_{n_r}$
MAD_1
MAD_2
....
MAD_r
$MMAD$
C
C^1
C^2
$\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3$
$\mathbf{y}^*$
$\mathbf{y}_1,\mathbf{y}_3$
$\mathbf{y}_2,\mathbf{y}^*$
TC^1
TC^2
$\mathbf{y}_2,\mathbf{y}_3$
$\mathbf{y}_1,\mathbf{y}^*$
$H_{(1)}$
$\mathbf{y}_1,\mathbf{y}_3$
$\mathbf{y}_2,\mathbf{y}^*$
$H_{(2)}$
$MH = H_{(2)} > 0$
$\mathbf{y}_1,\mathbf{y}_2$
$\mathbf{y}_3,\mathbf{y}^*$
$H_{(3)}$

Figure 3.7: The flow chart for the polythetic algorithm.

Let $MMAD$ be the maximum value of MAD_u for all clusters, $C_u,\ u = 1, \ldots, r$, in P_r. Then, $MMAD$ is defined as

$$MMAD = \max_u \{MAD_u,\ u = 1, \ldots, r\}. \tag{3.44}$$

- **Step 4**: Let $\mathbf{y}^*$ be the object with $MMAD$ value in P_r, and let $\mathbf{y}^* \in C_u$. Then, this object $\mathbf{y}^*$ is used as the seed for the bipartition of C_u, and the cluster C_u is bisected into (C_u^1, C_u^2). First of all, set $C_u^1 = \{\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(t)}\}$ and $C_u^2 = \{\mathbf{y}^*\}$, where $t = n_u - 1$. The C_u^1 is called the main cluster and C_u^2 with $\mathbf{y}^*$ is called the splinter cluster.

- **Step 5**: Set $TC^1 = C_u^1$ and $TC^2 = C_u^2$, and let one of the objects in TC^1, $\mathbf{y}_{(i)}$, move into TC^2. That is, now $TC^1 = TC^1 - \{\mathbf{y}_{(i)}\}$ and $TC^2 = TC^2 \cup \{\mathbf{y}_{(i)}\}$. Then, calculate the difference of the sums of the within-cluster variances between (C_u^1, C_u^2) and (TC^1, TC^2) that results from moving $\mathbf{y}_{(i)}$. That is, this difference, $H_{(i)}$, is

$$H_{(i)} = \{I(C_u^1) + I(C_u^2)\} - \{I(TC^1) + I(TC^2)\}. \tag{3.45}$$

  After $H_{(i)}$ for moving $\mathbf{y}_{(i)}$ is calculated, $\mathbf{y}_{(i)}$ goes back to TC^1. That is, now $TC^1 = TC^1 \cup \{\mathbf{y}_{(i)}\}$ and $TC^2 = TC^2 - \{\mathbf{y}_{(i)}\}$ (i.e., $TC^1 = C_u^1$ and $TC^2 = C_u^2$).

- **Step 6**: Repeat Step 5 for all objects in TC^1 (i.e., for all $\mathbf{y}_{(i)},\ i = 1, \ldots, t$). Then, we have t $H_{(i)}$ values corresponding to each $\mathbf{y}_{(i)},\ i = 1, \ldots, t$.

- **Step 7**: Let $MH = \max_i \{H_{(i)},\ i = 1, \ldots, t\}$ and $\mathbf{y}^{MH}$ be the object corresponding to the MH value in TC^1. Then, if $MH > 0$, the object $\mathbf{y}^{MH}$ moves from C_u^1 into C_u^2 (i.e., $C_u^1 = C_u^1 - \{\mathbf{y}^{MH}\}$ and $C_u^2 = C_u^2 \cup \{\mathbf{y}^{MH}\}$); and set $t = t - 1$.

- **Step 8**: Repeat Steps 5–7 while $MH > 0$. If $MH \leq 0$, go to the $(r+1)^{th}$ stage (i.e, $r = r + 1$).

- **Step 9**: Repeat Steps 2–8 until $r = R$ or $r = n$, where R is a prespecified number of clusters and n is the number of objects.

In Step 7, $MH > 0$ indicates that the difference of the sums of the within-cluster variances for moving of the object corresponding to the MH value is also positive because $MH = \max_i\{H_{(i)},\ i = 1, \ldots, t\}$. For example, suppose that the object $\mathbf{y}_{(2)}$ in the cluster C_u has the MH value and MH is positive. Then, the difference of the sums of the within-cluster variance for $\mathbf{y}_{(2)}$, $H_{(2)}$, is also positive. This means that the sum of within-cluster variances is decreased due to moving the object $\mathbf{y}_{(2)}$ into the splinter cluster. This coincides with our clustering criterion to minimize the total within-cluster variance of Equation (3.36). Thus, if MH is larger than zero, the object corresponding to the MH value moves into the splinter cluster. In contrast, if the MH value is negative, this means that the sum of within-cluster variances is increased due to moving the object with the MH value into the splinter cluster. In this case, that object stays in the main cluster. Figure 3.7 shows the process of the polythetic algorithm.

Example 3.6 *Consider the Ruspini (1970) histogram-valued data shown in Table 3.4 again. We illustrate the polythetic method using this dataset. At the first stage, we have* $\Omega \equiv P_1 = \{C_1\} = \big\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\big\}$. *First of all, we find a seed using the CDF dissimilarity matrix in Equation (3.41) and average weighted dissimilarity in Equation (3.42). If we use weights* $w_i = 1/n$ *in Equation (3.42), then the average weighted dissimilarity for object* $\mathbf{y}_1$ *is*

$$\bar{D}_u(\mathbf{y}_i) = \bar{D}_1(\mathbf{y}_1) = \frac{1}{4(4-1)}\big(104.946 + 127.868 + 94.750\big) = 27.297.$$

Similarly, $\bar{D}_1(\mathbf{y}_2) = 28.470$, $\bar{D}_1(\mathbf{y}_3) = 28.010$, *and* $\bar{D}_1(\mathbf{y}_4) = 30.924$. *Thus, since there is only one cluster at the first stage* ($P_1 = C_1$), *from Equation (3.43) and (3.44),* $MMAD = MAD_1 = 30.924$ *and the seed* $\mathbf{y}^* = \mathbf{y}_4$. *Therefore, the seed* $\mathbf{y}_4$ *automatically goes into the splinter cluster* C_1^2, *and we currently have the bipartition* $(C_1^1, C_1^2) = \big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_4\}\big)$.

Now, in turn, each object in the cluster C_1^1 *moves into the splinter cluster* C_1^2, *one at a time; and then we calculate the difference of the sums of the within-cluster variance in Equation (3.45). For example, suppose the object* $\mathbf{y}_{(1)} \equiv \mathbf{y}_1$ *moves into the splinter group; then the temporary bipartition* $(TC^1, TC^2) = \big(\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_4\}\big)$ *and the current bipartition*

$(C_1^1, C_1^2) = \big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_4\}\big)$. *From these bipartitions, the the difference of the sums of the within-cluster variance for* $\mathbf{y}_{(1)} \equiv \mathbf{y}_1$, $H_{(1)}$, *can be calculated using Equation (3.45) as follows:*

$$\begin{aligned}
I(C_1^1) &= \frac{1}{4 \times 3}\big(104.946^2 + 127.868^2 + 84.303^2\big) = 2872.56, \\
I(C_1^2) &= 0, \\
I(TC^1) &= \frac{1}{4 \times 2}\big(84.303^2\big) = 888.38, \\
I(TC^2) &= \frac{1}{4 \times 2}\big(94.750^2\big) = 1122.20.
\end{aligned}$$

Thus,

$$H_{(1)} = (2872.56 + 0) - (888.38 + 1122.20) = 861.98.$$

Similarly, we can calculate the $H_{(i)}$ *values for the other objects and this result is shown in Table 3.9.*

Table 3.9: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$ and $C_1^2 = \{\mathbf{y}_4\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_1$	$(\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_4\})$	888.38	1122.20	**+861.98**
$\mathbf{y}_{(2)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_1, \mathbf{y}_3\}, \{\mathbf{y}_2, \mathbf{y}_4\})$	2043.77	2902.89	-2074.10
$\mathbf{y}_{(3)} \equiv \mathbf{y}_3$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	1376.70	1920.48	-424.62

From Table 3.9, we know the maximum $H_{(i)}$ *value,* MH *is 861.98. Since* $MH = 861.98 > 0$, *the current bipartition of cluster* C_1, (C_1^1, C_1^2), *is* $(\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_4\})$. *Again, one of the objects in cluster* C_1^1 *moves into the splinter cluster* C_1^2 *at a time, and then the* $H_{(i)}$ *values are calculated. This procedure is similar to that used in obtaining Table 3.9. The result is shown in Table 3.10.*

Table 3.10: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_2, \mathbf{y}_3\}$ and $C_1^2 = \{\mathbf{y}_1, \mathbf{y}_4\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_3\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_4\})$	0.00	3601.19	-1590.62
$\mathbf{y}_{(2)} \equiv \mathbf{y}_3$	$(\{\mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_1, \mathbf{y}_4\})$	0.00	3390.96	-1380.39

From Table 3.10, since the maximum $H_{(i)}$ value, MH is -1380.39 and the MH value is less than zero, there is no object to move into the splinter cluster C_1^2 in this step. Thus, the optimal bipartition at the first stage is $(C_1^1, C_1^2) = (\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_4\})$.

At the second stage, we have a partition $P_2 = (C_1, C_2) = \big(\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_4\}\big)$. Similarly to the first stage, we find the seed using the average weighted dissimilarity in Equation (3.42). The average weighted dissimilarity for object $\mathbf{y}_2$ is

$$\bar{D}_u(\mathbf{y}_i) = \bar{D}_1(\mathbf{y}_2) = \frac{1}{4(2-1)}\big(84.303\big) = 21.076.$$

Similarly, $D_1(\mathbf{y}_3) = 21.076$, $D_2(\mathbf{y}_1) = 23.688$, and $D_2(\mathbf{y}_4) = 23.688$. Thus, since $MAD_1 = 21.076$ and $MAD_2 = 23.688$, the $MMAD$ value for the current partition is 23.688. This means the cluster C_2 is bipartitioned at this step. Since the cluster C_2 has only two objects and either object automatically moves into the splinter cluster, the optimal bipartition at the second stage is $(C_2^1, C_2^2) = \big(\{\mathbf{y}_1\}, \{\mathbf{y}_4\}\big)$. Thus, at the third stage, we have a partition $P_3 = (C_1, C_2, C_3) = \big(\{\mathbf{y}_2, \mathbf{y}_3\}, \{\mathbf{y}_1\}, \{\mathbf{y}_4\}\big)$. Similarly, we can find the optimal bipartition and binary question for the third stage. The dendrogram for the complete clustering result is shown in Figure 3.8.

The polythetic method has at most N_r bipartitions, where

$$N_r = \{n_u^*(n_u^* - 1)\}/2 - 1, \tag{3.46}$$

where n_u^* is the number of objects in the cluster that includes the object $\mathbf{y}^*$ with $MMAD$ value at the r^{th} stage. Thus, we need at most N_r repetitions for the Steps 5–7 at each stage, and this means there are at most N_r possible bipartitions to be considered at the r^{th} stage.

As mentioned in Section 3.4.1, the monothetic method evaluates $\sum_{u=1}^{r} p(n_u - 1)$ possible bipartitions to find the optimal bipartition at the r^{th} stage. From this, we know the number of possible bipartitions for the monothetic method depends on the number of variables p. In contrast, N_r in Equation (3.46) does not depend on the number of variables p. Thus, as the number of variables increases, the number of possible bipartitions for the monothetic method

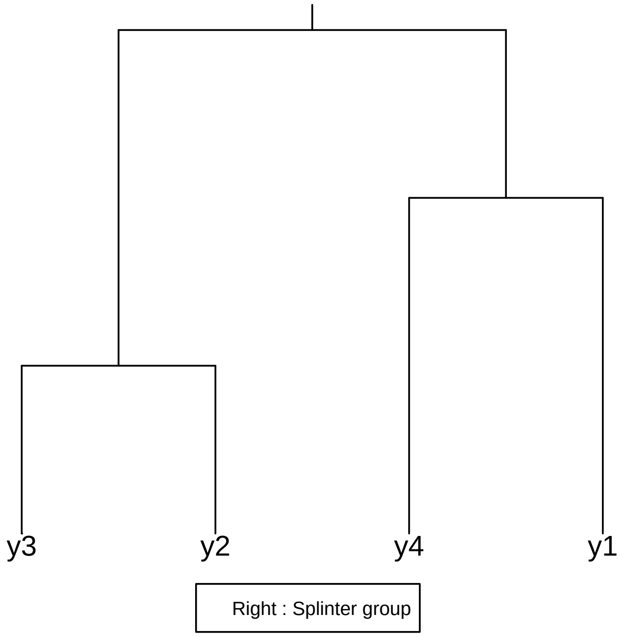


Figure 3.8: Dendrogram for Ruspini data using the polythetic method.

increases, but this number for the polythetic method does not change. In addition, to find the optimal bipartition, the monothetic method has to consider all $\sum_{u=1}^{r} p(n_u - 1)$ possible bipartitions at the r^{th} stage. In contrast, since the polythetic method stops the process for the r^{th} stage when the MH value is less than zero, the number of possible bipartitions for the polythetic method is, in practice, less than N_r. Therefore, when the number of variables is large, the polythetic method is more efficient in computing time than is the monothetic method. An application of this algorithm is shown in Chapter 5; and the program code in software R is provided in the Appendix C.

Chapter 4

Selection of the Optimal Number of Clusters

In any clustering approach, the entire set of objects is partitioned into r classes or clusters. To date, the literature does not provide details of how to find r in the clustering of symbolic objects. Thus, in these clustering methods, r is typically pre-defined, with the methodology applied to a range of possible r values. After providing some background in Section 4.1, we present in Section 4.2 some cluster validity indexes used to select an optimal r value. Then, in Section 4.3, these are evaluated and compared on some simulated data.

4.1 Background

The clustering methodology summarizes a collection of objects into a small number of classes by grouping objects with similar characteristics and separating objects with different characteristics. In general, we do not have any prior information about the number of clusters in the data. Kim and Ramakrishna (2005) indicated that the objects in each cluster are often indistinguishable under some criterion of similarity or dissimilarity. In addition, Everitt et al. (2001) indicated that different dissimilarity measures often lead to different groupings. Under this situation, we need to find the optimal number of clusters r and to evaluate clustering outcomes. Cluster validity indexes help to solve these problems.

Many different cluster validity indexes for classical data have been proposed in the literature, such as Dunn's (1974) index, Davis-Bouldin's (1979) index, and Xie-Beni's (1991) index, etc.; Milligan and Cooper (1985) has an extensive review of such indexes for classical data. The quality of the clustering outcomes depends on compactness and separability of clusters. Compactness means closeness of objects within a cluster, and separability indicates distinctness of between clusters. Berry and Linoff (1997) indicated that most cluster validity indexes are usually defined by combining these two properties. Compactness and separability are measured by within-cluster and between-cluster measures, respectively. Thus, small within-cluster and large between-cluster measures are deemed to give good clustering results.

Usually, since these indexes were developed for hierarchical clustering algorithms for classical or fuzzy datasets, they use the minimum (also called single linkage), maximum (or, complete linkage), and/or average dissimilarity values such as within or between-cluster measures. Since, unlike the case of classical data, the divisive clustering methods for symbolic objects proposed in Chapter 3 use the variance as within or between-cluster measures, existing cluster validity indexes would not be expected to work well for hierarchical divisive clustering algorithms for symbolic objects. Thus, cluster validity indexes for symbolic clustering should be different from those for classical data. Therefore, validity indexes for symbolic data such as interval-valued and histogram-valued data are developed in Section 4.2 and their properties studied through some simulation studies in Section 4.3.

4.2 Cluster Validity Indexes

Let $\mathbf{y}_i,\ i = 1, \ldots, n$, be symbolic objects with p-dimensional random variables $\{Y_j,\ j = 1, \ldots, p\}$ and $\mathbf{y}_i \in \Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. Also, let P_r be a partition of Ω at the r^{th} stage. Then, $P_r = \{C_u,\ u = 1, \ldots, r\}$, where $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$ is a cluster of size n_u, where $\sum_{u=1}^{r} n_u = n$. At the $(r + 1)^{th}$ stage, a single cluster C_u in P_r is bisected into C_u^1 and C_u^2. Thus, a new partition can be written as

$$P_{r+1} = \big(P_r \cup \{C_u^1, C_u^2\}\big) - \{C_u\}. \tag{4.1}$$

For a cluster $C_u = \{\mathbf{y}_1, \ldots, \mathbf{y}_{n_u}\}$, the within-cluster variance $I(C_u)$ is, for $u = 1, \ldots, r$,

$$I(C_u) = \frac{1}{2\tau} \sum_{i_1=1}^{n_u} \sum_{i_2=1}^{n_u} w_{i_1} w_{i_2} D^2(\mathbf{y}_{i_1}, \mathbf{y}_{i_2}),\ i_1, i_2 = 1, \ldots, n_u, \tag{4.2}$$

where $D(\mathbf{y}_{i_1}, \mathbf{y}_{i_2})$ is a dissimilarity or distance measure between the observation $\mathbf{y}_{i_1}$ and $\mathbf{y}_{i_2}$ in C_u and where w_i is the weight for the object $\mathbf{y}_i$, and where $\tau = \sum_{i=1}^{n_u} w_i$. For a partition $P_r = (C_1, \ldots, C_r)$ at the r^{th} stage, the total within-cluster variance is, for $r = 1, \ldots, n-1$,

$$W(P_r) = \sum_{u=1}^{r} I(C_u). \tag{4.3}$$

Also, the between-cluster variance is defined by, for $r = 1, \ldots, n-1$,

$$B(P_r) = W(\Omega) - W(P_r), \tag{4.4}$$

where $\Omega \equiv P_1$.

The divisive hierarchical clustering methods proposed in Chapter 3 use a within-cluster variance criterion. At each stage, our goal is to find the bipartition, $\left(C_u^1, C_u^2\right)$ of C_u, $u = 1, \ldots, r$, minimizing the total within-cluster variance, $W(P_r)$. Since, at the r^{th} stage, we have r clusters and one of the clusters is split into two clusters, we have $r+1$ clusters at the $(r+1)^{th}$ stage. Thus, as the procedure goes into the next stage, the number of clusters increases and the total within-cluster variance, $W(P_r)$, decreases. In addition, if the number of clusters is equal to the number of objects, then the total within-cluster variance is zero. That is, $W(P_{r+1}) \leq W(P_r)$ and $W(P_n) = 0$. In contrast, as the total within-cluster variance decreases, the between-cluster variance increases. Thus, the simplest index to find the optimal number of clusters would be the explained rate E_r given by, for $r = 1, \ldots, n$,

$$E_r = \frac{B(P_r)}{W(\Omega)} = \frac{W(\Omega) - W(P_r)}{W(\Omega)}, \tag{4.5}$$

where $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, and where $W(P_r)$ is the total within-cluster variance for the partition P_r defined in Equation (4.3) and $B(P_r)$ is the between-cluster variance for the partition P_r defined in Equation (4.4). The explained rate, E_r, is the ratio of the decrement of the within-cluster variance for the r^{th} stage and the total within-cluster variance of Ω, i.e., the proportion of the total variation explained by the variation between clusters. As the number of clusters increases, the explained rate increases because the total within-cluster variance,$W(P_r)$, decreases and the between-cluster variance, $B(P_r)$, increases. Let

$\delta_r = E_{r+1} - E_r$. Then, if δ_r suddenly becomes small relatively to δ_{r-1}, it indicates that the explained rate at the $(r+1)^{th}$ stage, E_{r+1}, has small increases relatively to E_r, and the effect of the bipartition from the r^{th} stage to the $(r+1)^{th}$ stage is weak. Thus, $r = r^*$ can be the optimal number of clusters.

However, through their simulation study, Milligan and Cooper (1985) showed that validity indexes such as the explained rate are not appropriate as a solution for determining the optimal number of clusters. They conducted a comparative evaluation of 30 validity indexes for classical data within a simulation framework and showed that such kinds of indexes using the decrement of the within-cluster inertia poorly perform relatively to other indexes. Chavent (1998) also refers to this fact. Thus, in this section we introduce two validity indexes for hierarchical divisive clustering for symbolic objects. These new indexes are based on Dunn's (1974) index and Davis-Bouldin's (1979) index developed for classical data.

For classical data, suppose that a partition P_r has r clusters (i.e., $P_r = (C_u,\ u = 1, \ldots, r)$) at the r^{th} stage. Then, the index introduced by Dunn (1974) is defined by

$$DI = \min_{u=1,\ldots,r} \left[\min_{t=1,\ldots,r,t\neq u} \left\{ \frac{d(C_t, C_u)}{\max_{l=1,\ldots,r}\{w(C_l)\}} \right\} \right], \tag{4.6}$$

where

$$d(C_t, C_u) = \min\{D(\mathbf{y}_i, \mathbf{y}_j) \mid \mathbf{y}_i \in C_t,\ \mathbf{y}_j \in C_u\},$$

and

$$w(C_l) = \max\{D(\mathbf{y}_i, \mathbf{y}_j) \mid \mathbf{y}_i, \mathbf{y}_j \in C_l\},$$

where $D(\mathbf{y}_i, \mathbf{y}_j)$ is the distance measure between two classical objects $\mathbf{y}_i$ and $\mathbf{y}_j$. Dunn's index is a function of the ratio of the minimum distance $d(C_t, C_u)$ between two clusters C_t and C_u and the maximum diameter $w(C_l)$ of clusters $C_l,\ l = 1, \ldots, r$. That is, $d(C_t, C_u)$ indicates the minimum distance between two clusters as a between-cluster measure, and $w(C_l)$ shows the maximum diameter of clusters as a within-cluster measure. Since well-separated clusters have large distances between clusters and small diameters within clusters, the partition with the higher Dunn's index value implies there is a better clustering result.

Davis and Bouldin (1979) proposed a cluster validity index that measures the ratio of the average of similarity measures between each cluster. It is defined as

$$DB = \frac{1}{r}\sum_{u=1}^{r}\left[\max_{t=1,\ldots,r,t\neq u}\left\{\frac{s_t + s_u}{d_{tu}}\right\}\right], \tag{4.7}$$

where

$$d_{tu} = D(\mathbf{v}_t, \mathbf{v}_u),$$

and

$$s_u = \frac{1}{n_u}\sum_{i=1}^{n_u} D(\mathbf{y}_i, \mathbf{v}_u),$$

where $D(\mathbf{v}_t, \mathbf{v}_u)$ is the distance measure between $\mathbf{v}_t$ and $\mathbf{v}_u$, where $\mathbf{v}_t$ and $\mathbf{v}_u$ are the centroids of the cluster C_u and C_t respectively, and n_u is the number of objects in C_u. That is, d_{tu} is the distance between two centroids of C_t and C_u and plays a role as the between-cluster measure. In contrast, s_u represents the dispersion of each cluster and is used as the within-cluster measure. A lower Davis-Bouldin's index value means that the clusters are compact and well-separated.

As mentioned in Chapter 3, the divisive clustering algorithms for symbolic objects use the within-cluster variance as a clustering criterion. Thus, we propose two cluster validity indexes for symbolic objects that use within-cluster and between-cluster variances. They are similar in concept to the Dunn and Davis-Bouldin's indexes. The within-cluster variance measures the compactness of each cluster and the between-cluster variance represents the separability between two clusters.

Definition 4.1 *For a partition $P_r = (C_1, \ldots, C_r)$ at the r^{th} stage, the* ***Dunn index for symbolic objects****, DI_r^s, for the partition P_r is given by*

$$DI_r^s = \min_{u=1,\ldots,r}\left[\min_{t=1,\ldots,r,t\neq u}\left\{\frac{I(C_t \cup C_u) - I(C_t) - I(C_u)}{\max_{l=1,\ldots,r}\{I(C_l)\}}\right\}\right], \; r = 2, \ldots, n-1, \tag{4.8}$$

where $I(\cdot)$ is the within-cluster variance from Definition 3.13.

This index consists of the ratio of the minimum between-cluster variance for all possible combinations of two clusters in the partition P_r and the maximum within-cluster variance for all clusters in P_r. As such, the index has the same design principles as does the Dunn index of Equation (4.6) for classical data. The between-cluster variance, $I(C_t \cup C_u) - I(C_t) - I(C_u)$, measures how far apart two clusters are, and the term $I(C_l)$ represents how close the objects are in each cluster. Since a good clustering result has small within-cluster measures and large between-cluster measures, to obtain a good clustering outcome, the denominator of the Dunn index in Equation (4.8) should be large and the numerator should be small. Thus, a higher DI_r^s value means a better clustering outcome has occurred.

Now, we propose the Davis-Bouldin index for symbolic objects. This index is defined by the average of cluster evaluation for all clusters in a partition.

Definition 4.2 *For a partition $P_r = (C_1, \ldots, C_r)$ at the r^{th} stage, the* ***Davis-Bouldin index for symbolic objects****, DB_r^s, for the partition P_r is defined as*

$$DB_r^s = \frac{1}{r} \sum_{u=1}^{r} \left[\frac{\max_{t=1,\ldots,r,t\neq u}\{I(C_t) + I(C_u)\}}{\min_{l=1,\ldots,r,l\neq u}\{I(C_l \cup C_u) - I(C_l) - I(C_u)\}} \right], \; r = 2, \ldots, n-1, \qquad (4.9)$$

where $I(\cdot)$ is the within-cluster variance from Definition 3.13.

This index consists of the average of ratios for each cluster in P_r of the maximum sum of two within-cluster variances and the minimum between-cluster measure for two clusters. Unlike the Dunn index for symbolic objects in Equation (4.8), the Davis-Bouldin index in Equation (4.9) uses average values of ratios of the within-cluster and the between-cluster measures. In addition, the ratio in Equation (4.9) is a little bit different from the that of the original Davis-Bouldin index in Equation (4.7). The original index uses the maximum ratio of the within-cluster and the between-cluster measures for each cluster. In contrast, the index of Equation (4.9) obtains the ratio of the maximum within-cluster measure and the minimum between-cluster measure. Similarly to the original index, a lower value of the Davis-Bouldin index for symbolic objects implies better results for the clustering outcome.

We illustrate the procedure calculating the Dunn and Davis-Bouldin indexes for symbolic objects in detail in Example 4.1.

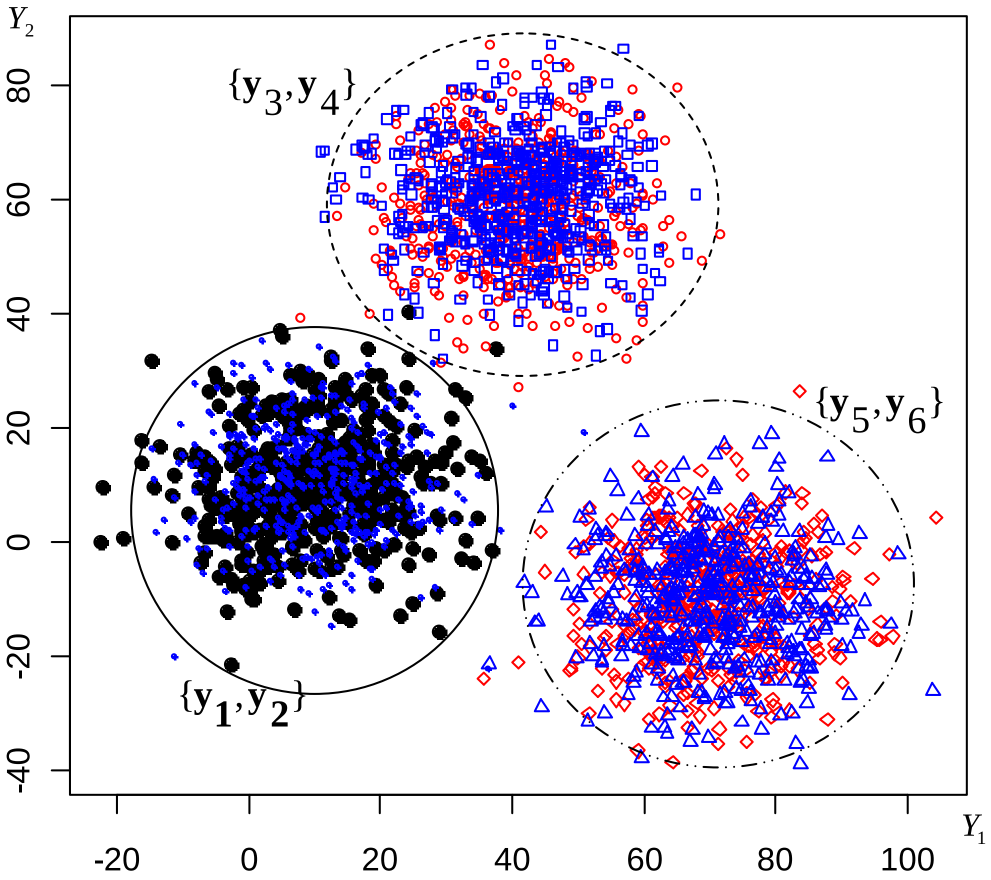


Figure 4.1: Struyf et al. (1997) data.

Example 4.1 *Consider the Struyf et al. (1997) classical dataset consisting of artificial data, with 3000 observations with two variables and three well-separated clusters as shown in Figure 4.1. The three clusters have 900, 1150, and 950 observations, respectively. We divide each cluster into two parts with equal size. For example, the first cluster is divided by two parts. One part has the* 1^{st} *to the* 450^{th} *observations, and the other part has the* 451^{st} *to the* 900^{th} *observations. Similarly, we can divide the second and third clusters by two parts, respectively. Then, we have a total of six parts (i.e., six objects) and generate six histogram-valued objects from the individual observations in each of the six parts. These histogram-valued data are* *shown in Table 4.1. From Table 4.1, transformed histogram-valued data can be obtained using Definition 3.3. These transformed histogram-valued data are shown in Table 4.2.*

Table 4.1: Struyf et al. (1997) histogram-valued data.

	Y_1	Y_2
$\mathbf{y}_1$	$\{[-30,-20),0.004;\ [-20,-10),0.022;\ [-10,0),0.164;\ [0,10),0.327;\ [10,20),0.320;\ [20,30),0.129;\ [30,40),0.033\}$	$\{[-30,-20),0.002;\ [-20,-10),0.018;\ [-10,0),0.129;\ [0,10),0.353;\ [10,20),0.318;\ [20,30),0.160;\ [30,40),0.018;\ [40,50),0.002\}$
$\mathbf{y}_2$	$\{[-20,-10),0.020;\ [-10,0),0.164;\ [0,10),0.331;\ [10,20),0.338;\ [20,30),0.127;\ [30,40),0.018;\ [40,50),0.000;\ [50,60),0.002\}$	$\{[-30,-20),0.004;\ [-20,-10),0.004;\ [-10,0),0.111;\ [0,10),0.327;\ [10,20),0.396;\ [20,30),0.131;\ [30,40),0.027\}$
$\mathbf{y}_3$	$\{[0,10),0.002;\ [10,20),0.016;\ [20,30),0.129;\ [30,40),0.325;\ [40,50),0.357;\ [50,60),0.153;\ [60,70),0.017;\ [70,80),0.002\}$	$\{[20,30),0.002;\ [30,40),0.031;\ [40,50),0.155;\ [50,60),0.348;\ [60,70),0.336;\ [70,80),0.113;\ [80,90),0.016\}$
$\mathbf{y}_4$	$\{[10,20),0.026;\ [20,30),0.146;\ [30,40),0.270;\ [40,50),0.369;\ [50,60),0.169;\ [60,70),0.021\}$	$\{[30,40),0.016;\ [40,50),0.106;\ [50,60),0.332;\ [60,70),0.410;\ [70,80),0.120;\ [80,90),0.016\}$
$\mathbf{y}_5$	$\{[30,40),0.002;\ [40,50),0.017;\ [50,60),0.131;\ [60,70),0.360;\ [70,80),0.337;\ [80,90),0.128;\ [90,100),0.023;\ [100,110),0.002\}$	$\{[-40,-30),0.023;\ [-30,-20),0.158;\ [-20,-10),0.314;\ [-10,0),0.339;\ [0,10),0.147;\ [10,20),0.017;\ [20,30),0.002\}$
$\mathbf{y}_6$	$\{[30,40),0.002;\ [40,50),0.023;\ [50,60),0.147;\ [60,70),0.343;\ [70,80),0.291;\ [80,90),0.166;\ [90,100),0.025;\ [100,110),0.002\}$	$\{[-40,-30),0.032;\ [-30,-20),0.143;\ [-20,-10),0.322;\ [-10,0),0.345;\ [0,10),0.128;\ [10,20),0.029\}$

Firstly, we illustrate the procedure calculating the Dunn and Davis-Bouldin indexes for symbolic objects using the extended Gowda-Diday measure and the polythetic algorithm; and then, we verify the Dunn and Davis-Bouldin indexes using the normalized city block distance and normalized Euclidean distance based on the extended Ichino-Yaguchi measure, and the normalized CDF distance, for both the monothetic and polythetic algorithms.

Table 4.2: Transformed histogram-valued data for Struyf et al. (1997) data.

Y_1	$\mathbf{y}_1'$	$\mathbf{y}_2'$	$\mathbf{y}_3'$	$\mathbf{y}_4'$	$\mathbf{y}_5'$	$\mathbf{y}_6'$
$[b_{1k}, b_{1,k+1})$	p_{11k}'	p_{21k}'	p_{31k}'	p_{41k}'	p_{51k}'	p_{61k}'
$[-30, -20)$	0.004	0	0	0	0	0
$[-20, -10)$	0.022	0.02	0	0	0	0
$[-10, 0)$	0.164	0.164	0	0	0	0
$[0, 10)$	0.327	0.331	0.002	0	0	0
$[10, 20)$	0.32	0.338	0.016	0.026	0	0
$[20, 30)$	0.129	0.127	0.129	0.146	0	0
$[30, 40)$	0.033	0.018	0.325	0.27	0.002	0.002
$[40, 50)$	0	0	0.357	0.369	0.017	0.023
$[50, 60)$	0	0.002	0.153	0.169	0.131	0.147
$[60, 70)$	0	0	0.017	0.021	0.36	0.343
$[70, 80)$	0	0	0.002	0	0.337	0.291
$[80, 90)$	0	0	0	0	0.128	0.166
$[90, 100)$	0	0	0	0	0.023	0.025
$[100, 110)$	0	0	0	0	0.002	0.002

Y_2	$\mathbf{y}_1'$	$\mathbf{y}_2'$	$\mathbf{y}_3'$	$\mathbf{y}_4'$	$\mathbf{y}_5'$	$\mathbf{y}_6'$
$[b_{2k}, b_{2,k+1})$	p_{12k}'	p_{22k}'	p_{32k}'	p_{42k}'	p_{52k}'	p_{62k}'
$[-40, -30)$	0	0	0	0	0.023	0.032
$[-30, -20)$	0.002	0.004	0	0	0.158	0.143
$[-20, -10)$	0.018	0.004	0	0	0.314	0.322
$[-10, 0)$	0.129	0.111	0	0	0.339	0.345
$[0, 10)$	0.353	0.327	0	0	0.147	0.128
$[10, 20)$	0.318	0.396	0	0	0.017	0.029
$[20, 30)$	0.16	0.131	0.002	0	0.002	0
$[30, 40)$	0.018	0.027	0.031	0.016	0	0
$[40, 50)$	0.002	0	0.155	0.106	0	0
$[50, 60)$	0	0	0.348	0.332	0	0
$[60, 70)$	0	0	0.336	0.41	0	0
$[70, 80)$	0	0	0.113	0.12	0	0
$[80, 90)$	0	0	0.016	0.016	0	0

From the transformed histogram-valued data of Table 4.2, the extended Gowda-Diday dissimilarity measure can be calculated by using Equation (3.21). Thus, the extended Gowda-Diday dissimilarity matrix for all variables of the transformed histogram-valued data of Table 4.2 is given by

$$\mathbf{D}_{GD} = \begin{pmatrix} 0 & 0.184 & 2.325 & 2.371 & 2.207 & 2.176 \\ 0.184 & 0 & 2.313 & 2.344 & 2.184 & 2.201 \\ 2.325 & 2.313 & 0 & 0.196 & 2.506 & 2.551 \\ 2.371 & 2.344 & 0.196 & 0 & 2.604 & 2.596 \\ 2.207 & 2.184 & 2.506 & 2.604 & 0 & 0.133 \\ 2.176 & 2.201 & 2.551 & 2.596 & 0.133 & 0 \end{pmatrix}. \tag{4.10}$$

Using the dissimilarity matrix in Equation (4.10), the polythetic method introduced in Section 3.4.2 can be performed for clustering the six objects, $\mathbf{y}_1, \ldots, \mathbf{y}_6$*. The clustering result is shown in Table 4.3 and the hierarchy in Figure 4.2. From Figure 4.2, we see that the cluster* $\{\mathbf{y}_1, \ldots, \mathbf{y}_6\}$ *is bipartitioned into* $\{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}$ *and* $\{\mathbf{y}_4, \mathbf{y}_3\}$ *at the first stage, and the splinter cluster is* $\{\mathbf{y}_4, \mathbf{y}_3\}$*. At the second stage, the cluster* $\{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}$ *is split into* $\{\mathbf{y}_6, \mathbf{y}_5\}$ *and* $\{\mathbf{y}_2, \mathbf{y}_1\}$*. Thus, the third partition,* P_3*, is* $\big(\{\mathbf{y}_6, \mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_4, \mathbf{y}_3\}\big)$*. This result coincides with the attribute of the classical dataset.*

Table 4.3: Clustering result using the polythetic method.

Partition P_r	Clusters $(C_1, \ldots, C_r)$
P_1	$\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}\big)$
P_2	$\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\},\ \{\mathbf{y}_3, \mathbf{y}_4\}\big)$
P_3	$\big(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3, \mathbf{y}_4\},\ \{\mathbf{y}_5, \mathbf{y}_6\}\big)$
P_4	$\big(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5, \mathbf{y}_6\}\big)$
P_5	$\big(\{\mathbf{y}_1\},\ \{\mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5, \mathbf{y}_6\}\big)$
P_6	$\big(\{\mathbf{y}_1\},\ \{\mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_6\}\big)$

Now, we can calculate the Dunn index for the clustering result of Table 4.3 using Equation (4.8). For the second partition P_2*, we have two clusters* $(C_1, C_2) = \big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\}\big)$*.*

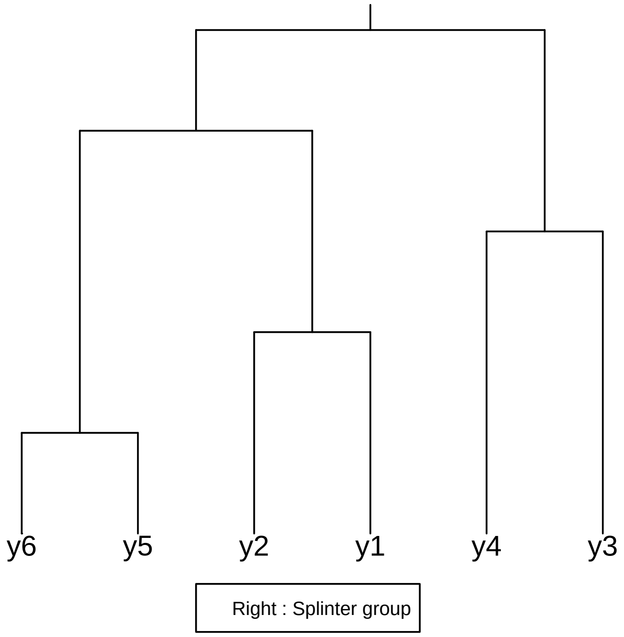


Figure 4.2: Dendrogram obtained by using the extended Gowda-Diday measure and the polythetic method.

The within-cluster variances for each cluster are given by, from Equation (3.35),

$$\begin{aligned} I(C_1) &= \frac{1}{6\times 4}\big(0.184^2+2.207^2+2.176^2+2.184^2+2.201^2+0.133^2\big) \\ &= 0.803, \qquad (4.11) \\ I(C_2) &= \frac{1}{6\times 2}\big(0.196^2\big)=0.003. \qquad (4.12) \end{aligned}$$

Thus,

$$\max_{l=1,2}\{I(C_l)\}=\max\{0.803,0.003\}=0.803. \qquad (4.13)$$

We calculate the between-cluster measure $I(C_t\cup C_u)-I(C_t)-I(C_u)$ as follows:

$$\begin{aligned} I(C_1\cup C_2) &= \frac{1}{6\times 6}\big(0.184^2+2.325^2+2.371^2+\cdots+2.596^2+0.133^2\big) \\ &= 1.875. \qquad (4.14) \end{aligned}$$

Thus,

$$I(C_1\cup C_2)-I(C_1)-I(C_2)=1.875-0.803-0.003=1.069. \qquad (4.15)$$

Therefore, the Dunn index for the partition at the second stage, DI_2^s is obtained as, from Equation (4.8),

$$DI_r^s = DI_2^s = \frac{1.069}{0.803} = 1.331. \tag{4.16}$$

For the partition at the third stage P_3, we have three clusters $(C_1, C_2, C_3) = \big(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\}\big)$. The Dunn index for this partition is obtained as follows: From Equation (3.35) and (4.10),

$$\begin{aligned}
I(C_1) &= \frac{1}{6 \times 2}\big(0.184^2\big) = 0.003, & (4.17)\\
I(C_2) &= \frac{1}{6 \times 2}\big(0.196^2\big) = 0.003, & (4.18)\\
I(C_3) &= \frac{1}{6 \times 2}\big(0.133^2\big) = 0.001. & (4.19)
\end{aligned}$$

Thus,

$$\max_{l=1,2,3}\{I(C_l)\} = \max\{0.003, 0.003, 0.001\} = 0.003. \tag{4.20}$$

Then, we consider all possible sets of union of any two clusters in the partition P_3, $C_t \cup C_u$, $t, u = 1, 2, 3$, $t \neq u$. The within-cluster variances for these sets of union are, from Equation (3.35) and (4.10),

$$\begin{aligned}
I(C_1 \cup C_2) &= I(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\})\\
&= \frac{1}{6 \times 4}\big(0.184^2 + 2.325^2 + 2.371^2 + 2.313^2 + 2.344^2 + 0.196^2\big)\\
&= 0.914, & (4.21)\\
I(C_1 \cup C_3) &= I(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\})\\
&= \frac{1}{6 \times 4}\big(0.184^2 + 2.207^2 + 2.176^2 + 2.184^2 + 2.201^2 + 0.133^2\big)\\
&= 0.803, & (4.22)\\
I(C_2 \cup C_3) &= I(\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})\\
&= \frac{1}{6 \times 4}\big(0.196^2 + 2.506^2 + 2.551^2 + 2.604^2 + 2.596^2 + 0.133^2\big)\\
&= 1.098. & (4.23)
\end{aligned}$$

Thus, the between-cluster measures for all these possible sets of union are

$$\begin{aligned} I(C_1 \cup C_2) - I(C_1) - I(C_2) &= 0.914 - 0.003 - 0.003 = 0.908, \quad (4.24) \\ I(C_1 \cup C_3) - I(C_1) - I(C_3) &= 0.803 - 0.003 - 0.001 = 0.799, \quad (4.25) \\ I(C_2 \cup C_3) - I(C_1) - I(C_3) &= 1.098 - 0.003 - 0.001 = 1.094. \quad (4.26) \end{aligned}$$

Hence, from Equation (4.8), the Dunn index for symbolic objects for the partition P_3 *is*

$$\begin{aligned} DI_r^s = DI_3^s &= \min\left\{\frac{0.908}{0.003}, \frac{0.799}{0.003}, \frac{1.094}{0.003}\right\} \\ &= \min\{286.880,\ 251.881,\ 344.593\} \\ &= 251.881. \quad (4.27) \end{aligned}$$

Similarly, we can obtain the Dunn index values for the partitions P_4 *and* P_5*. The complete set of results, for* $r = 2, \ldots, 5$*, is shown in Table 4.4.*

Table 4.4: Dunn index values

# of clusters	$r = 2$	$r = 3$	$r = 4$	$r = 5$
DI_r^s	1.331	*251.881*	1.128	1.912

From Table 4.4, we know that the Dunn index for the partition P_3 *has the largest value. This means the optimal number of clusters is three. This result coincides with the attribute of the original dataset.*

Now, we illustrate the Davis-Bouldin index in Equation (4.9) using the extended Gowda-Diday dissimilarity matrix Equation (4.10) and the divisive polythetic clustering result in Table 4.3. From Table 4.3, we know the partition at the second stage $P_2 = (C_1, C_2) = \big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\}\big)$.

Since there are only two clusters ($r = 2$*) in the partition, and* $I(C_1) + I(C_2) = 0.806$ *and* $I(C_1 \cup C_2) - I(C_1) - I(C_2) = 1.069$ *from Equation (4.11), (4.12), and (4.15), the Davis-Bouldin index for the partition at the second stage,* DB_2^s*, is obtained, from Equation (4.9), as*

$$DB_r^s = DB_2^s = \frac{1}{2}\left\{\frac{\max\{0.806\}}{\min\{1.069\}} + \frac{\max\{0.806\}}{\min\{1.069\}}\right\} = 0.754. \quad (4.28)$$

For the partition at the third stage P_3*, we have three clusters* $\big(C_1 = \{\mathbf{y}_1, \mathbf{y}_2\}, C_2 = \{\mathbf{y}_3, \mathbf{y}_4\},\ C_3 = \{\mathbf{y}_5, \mathbf{y}_6\}\big)$*. The Davis-Bouldin index for this partition is obtained as follows: For the cluster* C_1*, from Equation (4.17), (4.18), and (4.19),*

$$\begin{aligned} I(C_1) + I(C_2) &= 0.003 + 0.003 = 0.006, &(4.29)\\ I(C_1) + I(C_3) &= 0.003 + 0.001 = 0.004. &(4.30)\end{aligned}$$

Thus, the maximum within-cluster measure for C_1 *(i.e., the denominator of Equation (4.9) when* $u = 1$*) is*

$$\max_{t=2,3} \big\{I(C_t) + I(C_1)\big\} = \max\big\{0.006,\ 0.004\big\} = 0.006. \qquad (4.31)$$

Similarly, the maximum within-cluster measures for C_2 *and* C_3 *are* $\max_{t=1,3}\big\{I(C_t) + I(C_2)\big\} = 0.006$ *and* $\max_{t=1,2}\big\{I(C_t) + I(C_3)\big\} = 0.004$*, respectively.*

The minimum between-cluster measure for C_1 *(i.e., the numerator of Equation (4.9) when* $u = 1$*) is, from Equation (4.24) and (4.25),*

$$\min_{l=2,3} \big\{I(C_l \cup C_1) - I(C_l) - I(C_1)\big\} = \min\big\{0.908,\ 0.799\big\} = 0.799. \qquad (4.32)$$

Similarly, the minimum between-cluster measures for C_2 *and* C_3 *are* $\min_{l=1,3}\big\{I(C_l \cup C_2) - I(C_l) - I(C_2)\big\} = 0.908$ *and* $\min_{l=1,2}\big\{I(C_l \cup C_3) - I(C_l) - I(C_3)\big\} = 0.799$*, respectively.*

Hence, from Equation (4.9), the Davis-Bouldin index for symbolic objects for the partition P_3*,* DB^s_3*, is*

$$DB^s_r = DB^s_3 = \frac{1}{3}\left\{\frac{0.006}{0.799} + \frac{0.006}{0.908} + \frac{0.004}{0.799}\right\} = 0.006. \qquad (4.33)$$

Similarly, we can obtain the Davis-Bouldin index values for the partitions P_4 *and* P_5*. The complete set of results, for* $r = 2, \ldots, 5$*, is shown in Table 4.5.*

Table 4.5: Davis-Bouldin index values

# of clusters	$r = 2$	$r = 3$	$r = 4$	$r = 5$
DB^s_r	0.754	*0.006*	0.447	0.395

From Table 4.5, we know that the Davis-Bouldin index for the partition P_3 *has the smallest value. This means the optimal number of clusters is three. This identification of the optimal* $r = 3$ *is the same as the result for the Dunn index.*

Now, we verify the performance of cluster validity indexes using the city block distance of Equation (3.26) and Euclidean distances of Equation (3.27) based on the normalized extended Ichino-Yaguchi measure for $\gamma = 0.25$ defined in Equation (3.22) and (3.23), and the normalized CDF measure of Equation (3.31). Figure 4.1 shows that this dataset has obviously three clusters. Thus, for all distance or dissimilarity measures and clustering methods, the Dunn and Davis-Bouldin indexes should detect that the optimal number of clusters in this dataset is three.

From the transformed histogram-valued data of Table 4.2, we can calculate the normalized city block distance, the normalized Euclidean distance, and the normalized CDF measure.

The normalized city block distance is the normalized Minkowski distance of Equation (3.28) when q = 1, where, in this example, this measure is obtained by using the normalized extended Ichino-Yaguchi measure with $\gamma = 0.25$. Thus, we first calculate the normalized extended Ichino-Yaguchi measure for each variable. The normalized extended Ichino-Yaguchi dissimilarity matrix with $\gamma = 0.25$ for variable Y_1 is, from Equation (3.23),

$$\phi_j^* = \phi_1^* = \begin{pmatrix} 0 & 0.008 & 0.217 & 0.218 & 0.434 & 0.434 \\ 0.008 & 0 & 0.216 & 0.217 & 0.431 & 0.432 \\ 0.217 & 0.216 & 0 & 0.009 & 0.206 & 0.207 \\ 0.218 & 0.217 & 0.009 & 0 & 0.207 & 0.208 \\ 0.434 & 0.431 & 0.206 & 0.207 & 0 & 0.007 \\ 0.434 & 0.432 & 0.207 & 0.208 & 0.007 & 0 \end{pmatrix}. \tag{4.34}$$

The normalized extended Ichino-Yaguchi dissimilarity matrix with $\gamma = 0.25$ for variable Y_2 is, from Equation (3.23),

$$\phi_j^* = \phi_2^* = \begin{pmatrix} 0 & 0.014 & 0.374 & 0.388 & 0.153 & 0.152 \\ 0.014 & 0 & 0.369 & 0.384 & 0.158 & 0.158 \\ 0.374 & 0.369 & 0 & 0.012 & 0.537 & 0.538 \\ 0.388 & 0.384 & 0.012 & 0 & 0.552 & 0.553 \\ 0.153 & 0.158 & 0.537 & 0.552 & 0 & 0.010 \\ 0.152 & 0.158 & 0.538 & 0.553 & 0.010 & 0 \end{pmatrix}. \tag{4.35}$$

The sum of Equation (4.34) and (4.35) is the normalized city block distance matrix. Thus, the normalized city block distance matrix is given by

$$\mathbf{D}^1_{NM} = \begin{pmatrix} 0 & 0.022 & 0.591 & 0.606 & 0.586 & 0.587 \\ 0.022 & 0 & 0.585 & 0.601 & 0.589 & 0.589 \\ 0.591 & 0.585 & 0 & 0.021 & 0.744 & 0.745 \\ 0.606 & 0.601 & 0.021 & 0 & 0.759 & 0.760 \\ 0.586 & 0.589 & 0.744 & 0.759 & 0 & 0.017 \\ 0.587 & 0.589 & 0.745 & 0.760 & 0.017 & 0 \end{pmatrix}. \tag{4.36}$$

The normalized Euclidean distance is the normalized Minkowski distance of Equation (3.28) when $q = 2$, where, in this example, this measure is obtained by using the normalized extended Ichino-Yaguchi dissimilarity matrices with $\gamma = 0.25$ for each variable as given in Equation (4.34) and (4.35). From Equation (3.27) and (3.28), the normalized Euclidean distance matrix is

$$\mathbf{D}^2_{NM} = \begin{pmatrix} 0 & 0.016 & 0.433 & 0.445 & 0.460 & 0.460 \\ 0.016 & 0 & 0.427 & 0.441 & 0.459 & 0.459 \\ 0.433 & 0.427 & 0 & 0.015 & 0.576 & 0.577 \\ 0.445 & 0.441 & 0.015 & 0 & 0.590 & 0.590 \\ 0.460 & 0.459 & 0.576 & 0.590 & 0 & 0.012 \\ 0.460 & 0.459 & 0.577 & 0.590 & 0.012 & 0 \end{pmatrix}. \tag{4.37}$$

The normalized CDF measure is obtained by using Equation (3.31). Thus, the normalized CDF dissimilarity matrix is

$$\mathbf{D}_{NCDF} = \begin{pmatrix} 0 & 0.014 & 0.595 & 0.609 & 0.589 & 0.590 \\ 0.014 & 0 & 0.590 & 0.604 & 0.595 & 0.596 \\ 0.595 & 0.590 & 0 & 0.020 & 0.741 & 0.742 \\ 0.609 & 0.604 & 0.020 & 0 & 0.753 & 0.754 \\ 0.589 & 0.595 & 0.741 & 0.753 & 0 & 0.008 \\ 0.590 & 0.596 & 0.742 & 0.754 & 0.008 & 0 \end{pmatrix}. \tag{4.38}$$

Using these distance or dissimilarity matrices, we can perform the divisive monothetic and polythetic clustering methods. The clustering results for each distance or dissimilarity measure and clustering method are shown in Table 4.6.

Table 4.6: Clustering results for each measure and method.

		Monothetic method	Polythetic method
Measure	P_r	Clusters $(C_1, \ldots, C_r)$	Clusters $(C_1, \ldots, C_r)$
Extended Gowda-Diday	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_6	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$
Normalized city block	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_4	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_6	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$
Normalized Euclidean	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_4	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_6	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$
Normalized CDF	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\})$
	P_6	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\}, \{\mathbf{y}_6\})$

We see from Table 4.6 that both the monothetic and polythetic methods give the same clustering results for each given distance/dissimilarity measure. In contrast, the clustering results for each dissimilarity measure are a little bit different depending on which measure was used in the clustering method. However, the partitions at the third stage are all the same regardless of measures and methods (i.e., $P_3 = \big(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_6\}\big)$). This partition coincides with the attribute of the original classical dataset. From the distance or dissimilarity *matrices in Equation (4.36), (4.37), and (4.38) and clustering outcomes in Table 4.6, we can calculate the Dunn and Davis-Bouldin indexes using Equation (4.8) and (4.9), respectively.*

Table 4.7: Cluster validity index values for Example 4.1

		Monothetic method				Polythetic method			
Measure	Validity	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 2$	$r = 3$	$r = 4$	$r = 5$
Extended	DI_r^s	1.331	*250.531*	1.128	1.912	1.331	*250.531*	1.128	1.912
Gowda-Diday	DB_r^s	0.754	*0.007*	0.447	0.395	0.754	*0.007*	0.447	0.395
Normalized	DI_r^s	1.442	*1389.382*	1.157	1.477	1.442	*1389.382*	1.157	1.477
city block	DB_r^s	0.694	*0.001*	0.433	0.505	0.694	*0.001*	0.433	0.505
Normalized	DI_r^s	1.593	*1440.352*	1.204	1.486	1.593	*1440.352*	1.204	1.486
Euclidean	DB_r^s	0.628	*0.001*	0.416	0.493	0.628	*0.001*	0.416	0.493
Normalized	DI_r^s	1.411	*1827.749*	2.037	2.620	1.411	*1827.749*	2.037	2.620
CDF	DB_r^s	0.709	*0.001*	0.246	0.228	0.709	*0.001*	0.246	0.228

Table 4.7 shows the Dunn and Davis-Bouldin index values for the extended Gowda-Diday measure, the normalized city block distance and normalized Euclidean distance based on the normalized extended Ichino-Yaguchi measure, and the normalized CDF distance, for both the divisive monothetic and polythetic clustering methods. For all measures and methods, when the number of clusters are three, the Dunn index values, DI_r^s, are relatively very large and the Davis-Bouldin index values, DB_r^s, are relatively very small. This means that the optimal number of clusters is clearly identified as three.

4.3 Simulation Study

In this section, we evaluate the performance of the Dunn index and Davis-Bouldin index for symbolic objects by a simulation study through three examples. For the simulation, the true number of clusters for datasets is known. We also consider various dissimilarity measures such as the extended Gowda-Diday measure, the normalized city block distance and normalized Euclidean distance based on the normalized extended Ichino-Yaguchi measure, and the normalized CDF measure, for both the monothetic and polythetic mehtods. To verify the cluster validity indexes proposed in Section 4.2, in Example 4.2, simulated histogram-valued data are used; also, the cluster validity indexes are verified through Fisher's (1936) iris dataset in Example 4.3.

Example 4.2 *We use the dataset of Table 4.8 shown in Figure 4.3 to verify that the cluster validity indexes proposed in Section 4.2 can work well for the case of clustering for histogram-valued data generated by bivariate normal random numbers. To obtain simulated histogram-valued data with two variables, we generate random numbers from bivariate normal distributions as follows:*

$$\mathbf{y}_1 \textit{ and } \mathbf{y}_2 \sim N_2\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}\right), \tag{4.39}$$

$$\mathbf{y}_3 \textit{ and } \mathbf{y}_4 \sim N_2\left(\begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{pmatrix}\right), \tag{4.40}$$

$$\mathbf{y}_5 \sim N_2\left(\begin{pmatrix} 10 \\ 5 \end{pmatrix}, \begin{pmatrix} 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}\right). \tag{4.41}$$

Thus, 200 classical sample points for each object are generated from the bivariate normal distributions, and then histogram-valued data are found from these generated classical sample points. For example, to obtain the histogram-valued data for the object $\mathbf{y}_1$, we sample 200 classical points from the bivariate normal distribution of Equation (4.39), and then we find histogram-valued data for object $\mathbf{y}_1$ from these 200 sample points. These histogram-valued data for all objects are shown in Table 4.8. Plots of the individual classical data points are shown in Figure 4.3.

As shown in Figure 4.3, there are three clusters, $(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$, in the dataset. From Equation (4.39) and (4.40), the clusters $\{\mathbf{y}_1, \mathbf{y}_2\}$ and $\{\mathbf{y}_3, \mathbf{y}_4\}$ are generated from the same mean vector and different variance-covariance matrices. In contrast, from Equation and (4.39) and and (4.41), the clusters $\{\mathbf{y}_1, \mathbf{y}_2\}$ and $\{\mathbf{y}_5\}$ have the same variance-covariance matrix, but the mean vectors are different. Comparing Figure 4.1 and Figure 4.3, it is clear

Table 4.8: Simulated histogram-valued data for Example 4.2.

	Y_1	Y_2
$\mathbf{y}_1$	$\{[3.5, 4.0), 0.020;\ [4.0, 4.5), 0.165;\ [4.5, 5.0), 0.325;\ [5.0, 5.5), 0.290;\ [5.5, 6.0), 0.150;\ [6.0, 6.5), 0.050\}$	$\{[3.5, 4.0), 0.035;\ [4.0, 4.5), 0.160;\ [4.5, 5.0), 0.345;\ [5.0, 5.5), 0.300;\ [5.5, 6.0), 0.120;\ [6.0, 6.5), 0.025;\ [6.5, 7.0), 0.015\}$
$\mathbf{y}_2$	$\{[3.0, 3.5), 0.005;\ [3.5, 4.0), 0.020;\ [4.0, 4.5), 0.160;\ [4.5, 5.0), 0.300;\ [5.0, 5.5), 0.340;\ [5.5, 6.0), 0.150;\ [6.0, 6.5), 0.020;\ [6.5, 7.0), 0.005\}$	$\{[3.5, 4.0), 0.035;\ [4.0, 4.5), 0.165;\ [4.5, 5.0), 0.275;\ [5.0, 5.5), 0.305;\ [5.5, 6.0), 0.165;\ [6.0, 6.5), 0.050;\ [6.5, 7.0), 0.005\}$
$\mathbf{y}_3$	$\{[2.0, 3.0), 0.025;\ [3.0, 4.0), 0.135;\ [4.0, 5.0), 0.360;\ [5.0, 6.0), 0.345;\ [6.0, 7.0), 0.125;\ [7.0, 8.0), 0.010; \}$	$\{[2.0, 3.0), 0.020;\ [3.0, 4.0), 0.155;\ [4.0, 5.0), 0.295;\ [5.0, 6.0), 0.345;\ [6.0, 7.0), 0.170;\ [7.0, 8.0), 0.015\}$
$\mathbf{y}_4$	$\{[2.0, 3.0), 0.050;\ [3.0, 4.0), 0.130;\ [4.0, 5.0), 0.355;\ [5.0, 6.0), 0.290;\ [6.0, 7.0), 0.145;\ [7.0, 8.0), 0.030; \}$	$\{[2.0, 3.0), 0.030;\ [3.0, 4.0), 0.135;\ [4.0, 5.0), 0.360;\ [5.0, 6.0), 0.330;\ [6.0, 7.0), 0.115;\ [7.0, 8.0), 0.025;\ [8.0, 9.0), 0.005\}$
$\mathbf{y}_5$	$\{[8.0, 8.5), 0.005;\ [8.5, 9.0), 0.020;\ [9.0, 9.5), 0.205;\ [9.5, 10.0), 0.255;\ [10.0, 10.5), 0.340;\ [10.5, 11.0), 0.155;\ [11.0, 11.5), 0.015;\ [11.5, 12.0), 0.005\}$	$\{[3.0, 3.5), 0.005;\ [3.5, 4.0), 0.020;\ [4.0, 4.5), 0.150;\ [4.5, 5.0), 0.365;\ [5.0, 5.5), 0.300;\ [5.5, 6.0), 0.135;\ [6.0, 6.5), 0.020;\ [6.5, 7.0), 0.005\}$

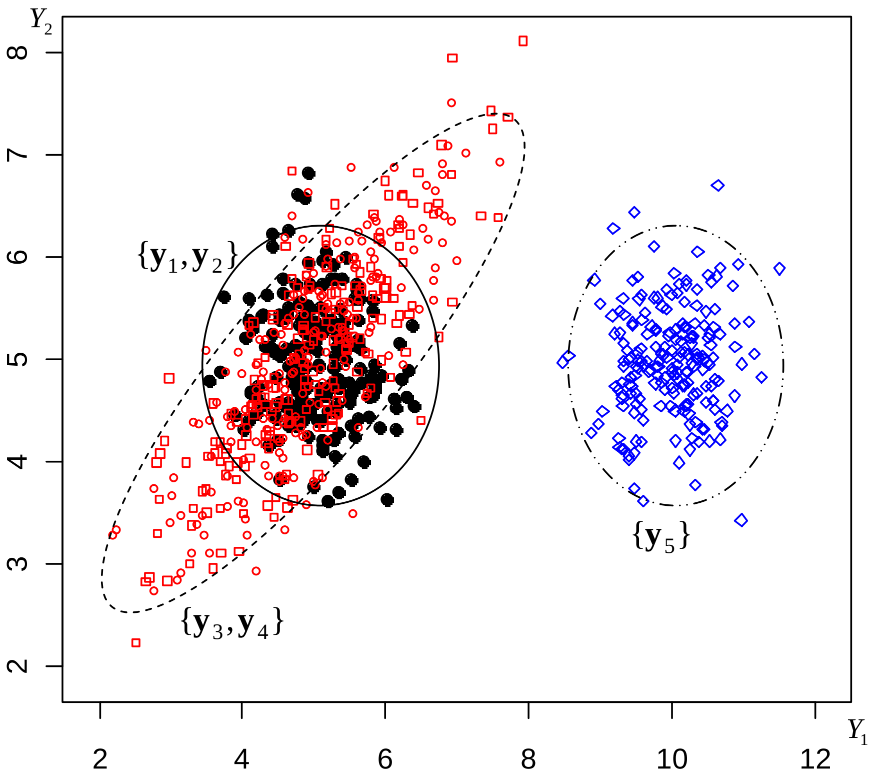


Figure 4.3: Simulated data for Example 4.2.

that the data in this example are different from the data of Example 4.1. The three clusters in Example 4.1 are obviously distinguishable because they do not overlap. In contrast, the clusters $\{\mathbf{y}_1, \mathbf{y}_2\}$ *and* $\{\mathbf{y}_3, \mathbf{y}_4\}$ *in the present example do overlap. Thus, in this example, the cluster validity indexes are verified for the case where clusters overlap.*

The transformed histogram-valued data for the data of Table 4.8 are obtained using Definition 3.3 and are shown in Table 4.9. From these transformed histogram-valued data of Table 4.9, we can calculate the extended Gowda-Diday measure of Equation (3.21), the city block distance of Equation (3.26) and Euclidean distances of Equation (3.27) based on the normalized extended Ichino-Yaguchi measure for $\gamma = 0.25$ *defined in Equation (3.22) and (3.23), and the normalized CDF measure of Equation (3.31).*

Table 4.9: Transformed histogram-valued data for Example 4.2.

Y_1	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$
$[b_{1k}, b_{1,k+1})$	p'_{11k}	p'_{21k}	p'_{31k}	p'_{41k}	p'_{51k}
$[2, 2.5)$	0	0	0.012	0.025	0
$[2.5, 3)$	0	0	0.012	0.025	0
$[3, 3.5)$	0	0.005	0.068	0.065	0
$[3.5, 4)$	0.02	0.02	0.068	0.065	0
$[4, 4.5)$	0.165	0.16	0.18	0.178	0
$[4.5, 5)$	0.325	0.3	0.18	0.178	0
$[5, 5.5)$	0.29	0.34	0.172	0.145	0
$[5.5, 6)$	0.15	0.15	0.172	0.145	0
$[6, 6.5)$	0.05	0.02	0.062	0.072	0
$[6.5, 7)$	0	0.005	0.062	0.072	0
$[7, 7.5)$	0	0	0.005	0.015	0
$[7.5, 8)$	0	0	0.005	0.015	0
$[8, 8.5)$	0	0	0	0	0.005
$[8.5, 9)$	0	0	0	0	0.02
$[9, 9.5)$	0	0	0	0	0.205
$[9.5, 10)$	0	0	0	0	0.255
$[10, 10.5)$	0	0	0	0	0.34
$[10.5, 11)$	0	0	0	0	0.155
$[11, 11.5)$	0	0	0	0	0.015
$[11.5, 12)$	0	0	0	0	0.005
Y_2	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$
$[b_{2k}, b_{2,k+1})$	p'_{12k}	p'_{22k}	p'_{32k}	p'_{42k}	p'_{52k}
$[2, 2.5)$	0	0	0.01	0.015	0
$[2.5, 3)$	0	0	0.01	0.015	0
$[3, 3.5)$	0	0	0.078	0.068	0.005
$[3.5, 4)$	0.035	0.035	0.078	0.068	0.02
$[4, 4.5)$	0.16	0.165	0.148	0.18	0.15
$[4.5, 5)$	0.345	0.275	0.148	0.18	0.365
$[5, 5.5)$	0.3	0.305	0.172	0.165	0.3
$[5.5, 6)$	0.12	0.165	0.172	0.165	0.135
$[6, 6.5)$	0.025	0.05	0.085	0.058	0.02
$[6.5, 7)$	0.015	0.005	0.085	0.058	0.005
$[7, 7.5)$	0	0	0.008	0.012	0
$[7.5, 8)$	0	0	0.008	0.012	0
$[8, 8.5)$	0	0	0	0.002	0
$[8.5, 9)$	0	0	0	0.002	0

The extended Gowda-Diday dissimilarity matrix is given by, from Equation (3.8),

$$\mathbf{D}_{GD} = \begin{pmatrix} 0 & 0.185 & 1.205 & 1.316 & 1.600 \\ 0.185 & 0 & 1.180 & 1.304 & 1.659 \\ 1.205 & 1.180 & 0 & 0.241 & 2.460 \\ 1.316 & 1.304 & 0.241 & 0 & 2.517 \\ 1.600 & 1.659 & 2.460 & 2.517 & 0 \end{pmatrix}. \qquad (4.42)$$

The normalized city block distance is the normalized Minkowski distance of Equation (3.28) when $q = 1$, where, in this example, this measure is obtained by using the normalized extended Ichino-Yaguchi measure with $\gamma = 0.25$. Thus, we first calculate the normalized extended Ichino-Yaguchi measure for each variable. The normalized extended Ichino-Yaguchi dissimilarity matrix with $\gamma = 0.25$ for variable Y_1 is, from Equation (3.23),

$$\phi_j^* = \phi_1^* = \begin{pmatrix} 0 & 0.008 & 0.052 & 0.068 & 0.493 \\ 0.008 & 0 & 0.054 & 0.070 & 0.494 \\ 0.052 & 0.054 & 0 & 0.019 & 0.498 \\ 0.068 & 0.070 & 0.019 & 0 & 0.499 \\ 0.493 & 0.494 & 0.498 & 0.499 & 0 \end{pmatrix}. \qquad (4.43)$$

The normalized extended Ichino-Yaguchi dissimilarity matrix with $\gamma = 0.25$ for variable Y_2 is, from Equation (3.23),

$$\phi_j^* = \phi_2^* = \begin{pmatrix} 0 & 0.014 & 0.087 & 0.087 & 0.013 \\ 0.014 & 0 & 0.082 & 0.082 & 0.017 \\ 0.087 & 0.082 & 0 & 0.029 & 0.093 \\ 0.087 & 0.082 & 0.029 & 0 & 0.094 \\ 0.013 & 0.017 & 0.093 & 0.094 & 0 \end{pmatrix}. \qquad (4.44)$$

The sum of Equation (4.43) and (4.44) is the normalized city block distance matrix. Thus, the normalized city block distance matrix is given by

$$\mathbf{D}^1_{NM} = \begin{pmatrix} 0 & 0.022 & 0.139 & 0.156 & 0.506 \\ 0.022 & 0 & 0.137 & 0.152 & 0.511 \\ 0.139 & 0.137 & 0 & 0.047 & 0.591 \\ 0.156 & 0.152 & 0.047 & 0 & 0.593 \\ 0.506 & 0.511 & 0.591 & 0.5930 & \end{pmatrix}. \tag{4.45}$$

The normalized Euclidean distance is the normalized Minkowski distance of Equation (3.28) when $q = 2$, *where, in this example, this measure is obtained by using the normalized extended Ichino-Yaguchi dissimilarity matrices with* $\gamma = 0.25$ *for each variable as given in Equation (4.43) and (4.44). From Equation (3.27) and (3.28), the normalized Euclidean distance matrix is*

$$\mathbf{D}^2_{NM} = \begin{pmatrix} 0 & 0.016 & 0.102 & 0.111 & 0.493 \\ 0.016 & 0 & 0.099 & 0.108 & 0.495 \\ 0.102 & 0.099 & 0 & 0.034 & 0.507 \\ 0.111 & 0.108 & 0.034 & 0 & 0.508 \\ 0.493 & 0.495 & 0.507 & 0.508 & 0 \end{pmatrix}. \tag{4.46}$$

The normalized CDF measure is obtained by using Equation (3.31). Thus, the normalized CDF dissimilarity matrix is

$$\mathbf{D}_{NCDF} = \begin{pmatrix} 0 & 0.016 & 0.101 & 0.106 & 0.500 \\ 0.016 & 0 & 0.094 & 0.106 & 0.511 \\ 0.101 & 0.094 & 0 & 0.028 & 0.573 \\ 0.106 & 0.106 & 0.028 & 0 & 0.566 \\ 0.500 & 0.511 & 0.573 & 0.566 & 0 \end{pmatrix}. \tag{4.47}$$

Using these distance or dissimilarity matrices, we can perform the divisive monothetic and polythetic clustering methods. The clustering results for each distance or dissimilarity measure and clustering method are shown in Table 4.10.

Table 4.10: Clustering results for each measure and method.

		Monothetic method	Polythetic method
Measure	P_r	Clusters $(C_1, \ldots, C_r)$	Clusters $(C_1, \ldots, C_r)$
Extended Gowda-Diday	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
Normalized City block	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
Normalized Euclidean	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
Normalized CDF	P_1	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\})$
	P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_3	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_4	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$
	P_5	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2\}, \{\mathbf{y}_3\}, \{\mathbf{y}_4\}, \{\mathbf{y}_5\})$

From Table 4.10, we see that both the monothetic and polythetic methods give the same clustering results for any one given distance/dissimilarity measures, and the four distance/dissimilarity measures also give the same outcomes for both the monothetic and polythetic algorithms.

For the distance or dissimilarity matrices in Equation (4.42), (4.45), (4.46), and (4.47) and clustering outcomes in Table 4.10, we can calculate the Dunn and Davis-Bouldin indexes using Equation (4.8) and (4.9).

Table 4.11 shows the Dunn and Davis-Bouldin index values for the extended Gowda-Diday measure, the normalized city block distance and normalized Euclidean distance based on the normalized extended Ichino-Yaguchi measure, and the normalized CDF distance, for both

Table 4.11: Cluster validity index values for Example 4.2

		Monothetic method			Polythetic method		
Measure	Validity	$r=2$	$r=3$	$r=4$	$r=2$	$r=3$	$r=4$
Extended	DI_r^s	2.023	*53.387*	1.698	2.023	*53.387*	1.698
Gowda-Diday	DB_r^s	0.494	*0.025*	0.301	0.494	*0.025*	0.301
Normalized	DI_r^s	10.833	*18.487*	4.541	10.833	*18.487*	4.541
city block	DB_r^s	0.092	*0.046*	0.115	0.092	*0.046*	0.115
Normalized	DI_r^s	17.436	*18.197*	4.453	17.436	*18.197*	4.453
Euclidean	DB_r^s	0.057	*0.046*	0.117	0.057	*0.046*	0.117
Normalized	DI_r^s	21.600	*25.772*	3.242	21.600	*25.772*	3.242
CDF	DB_r^s	0.046	*0.035*	0.159	0.046	*0.035*	0.159

the divisive monothetic and polythetic clustering methods. For all measures and methods, when the number of clusters are three, the Dunn index values, DI_r^s*, are largest and the Davis-Bouldin index values,* DB_r^s*, are smallest. From Table 4.10, the three clusters are* $\big(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5\}\big)$*. However, index values at* $r = 3$ *are relatively close to those at* $r = 2$*. This can be explained by the fact that clusters* $\{\mathbf{y}_1, \mathbf{y}_2\}$ *and* $\{\mathbf{y}_3, \mathbf{y}_4\}$ *have the same mean vector and overlap. Nevertheless, the validity indexes detect the optimal number of clusters (three) in this example even with the overlap.*

Example 4.3 *In this example, we evaluate the proposed indexes for the Fisher's (1936) iris dataset used in Example 3.1. Originally, Fisher's iris dataset has four variables, viz.,* $Y_1 =$*'Sepal Length',* $Y_2 =$*'Sepal Width',* $Y_3 =$*'Petal Length', and* $Y_4 =$*'Petal Width' and 150 individual observations from three species (setosa, versicolor, virginica), with each species having 50 individual observations. However, in this example, we aggregate consecutive groups of ten from the 150 observations. After aggregating, a histogram-valued dataset with 15 objects can be generated (these are provided in the Appendix A.1, Table A.1). Each iris species is represented by five symbolic objects. For example, object* $\mathbf{y}_1, \dots, \mathbf{y}_5$ *are included in the species 'setosa',* $\mathbf{y}_6, \dots, \mathbf{y}_{10}$ *are in the species 'versicolor', and* $\mathbf{y}_{10}, \dots, \mathbf{y}_{15}$ *are in the species 'virginica'.*

It is well known that the Fisher's iris dataset has two or three clusters. Thus, the true number of clusters for the histogram-valued iris dataset is also two or three. If it has three clusters, each species constitutes one cluster. If two clusters, one cluster is the species 'setosa', and another cluster has the species 'versicolor' and 'virginica'.

The transformed histogram-valued data are obtained using Definition 3.3 and are given in Table A.2 in Appendix A.2. From these transformed histogram-valued data, we can calculate the extended Gowda-Diday measure of Equation (3.21), the city block distance of Equation (3.26) and Euclidean distances of Equation (3.27) based on the normalized extended Ichino-Yaguchi measure for $\gamma = 0.25$ *defined in Equation (3.22) and (3.23), and the normalized CDF measure of Equation (3.31).*

From Equation (3.8), the extended Gowda-Diday dissimilarity matrix is given in Table 4.12. The normalized city block distance is the normalized Minkowski distance of Equation (3.28) when $q = 1$*, where, in this example, this measure is obtained by using the normalized extended Ichino-Yaguchi measure with* $\gamma = 0.25$*. Thus, we first calculate the normalized extended Ichino-Yaguchi measure for each variable. The normalized extended Ichino-Yaguchi dissimilarity matrices with* $\gamma = 0.25$ *for each variable (i.e.,* $Y_1, \dots, Y_4$*) are, from Equation (3.23), given in Table 4.13, 4.14, 4.15, and 4.16, respectively. The sum of the normalized extended Ichino-Yaguchi dissimilarity matrices given in Table 4.13, 4.14, 4.15, and 4.16 is the normalized city block distance matrix. Thus, the normalized city block distance matrix is given in Table 4.17. The normalized Euclidean distance is the normalized Minkowski distance of Equation (3.28) when* $q = 2$*, where this measure is obtained by using the normalized extended Ichino-Yaguchi dissimilarity matrices of Equation (3.23) with* $\gamma = 0.25$ *as given in Table 4.13, 4.14, 4.15, and 4.16. From Equation (3.27) and (3.28), the normalized Euclidean distance matrix is shown in Table 4.18. The normalized CDF measure is obtained by using Equation (3.31). Thus, the normalized CDF dissimilarity matrix is shown in Table 4.19.*

Using these distance or dissimilarity matrices, we can perform the monothetic and polythetic divisive clustering methods. The clustering results for each of the extended Gowda-

Table 4.12: The extended Gowda-Diday dissimilarity matrix.

$\mathbf{D}_{GD}$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	2.34	3.14	1.89	2.68	6.36	6.51	6.73	6.39	6.21	7.25	7.42	7.46	7.34	6.71
$\mathbf{y}_2$	2.34	0	3.35	2.09	3.15	5.87	5.91	6.73	6.00	5.91	7.03	6.94	7.18	6.80	6.80
$\mathbf{y}_3$	3.14	3.35	0	3.55	3.06	6.53	6.63	6.48	6.58	6.02	7.66	7.68	7.27	7.11	6.66
$\mathbf{y}_4$	1.89	2.09	3.55	0	3.27	6.39	6.33	6.94	6.44	6.00	7.60	7.60	7.67	7.52	6.97
$\mathbf{y}_5$	2.68	3.15	3.06	3.27	0	5.98	5.89	6.48	5.97	5.85	7.18	6.67	7.42	7.03	6.53
$\mathbf{y}_6$	6.36	5.87	6.53	6.39	5.98	0	2.47	2.58	1.91	2.76	4.03	4.08	4.08	4.08	4.67
$\mathbf{y}_7$	6.51	5.91	6.63	6.33	5.89	2.47	0	2.52	1.90	2.83	4.93	4.42	4.78	4.86	4.29
$\mathbf{y}_8$	6.73	6.73	6.48	6.94	6.48	2.58	2.52	0	2.09	2.70	4.46	4.47	3.93	4.25	3.45
$\mathbf{y}_9$	6.39	6.00	6.58	6.44	5.97	1.91	1.90	2.09	0	3.08	4.55	4.36	4.54	4.40	3.98
$\mathbf{y}_{10}$	6.21	5.91	6.02	6.00	5.85	2.76	2.83	2.70	3.08	0	5.27	5.42	5.12	5.55	5.08
$\mathbf{y}_{11}$	7.25	7.03	7.66	7.60	7.18	4.03	4.93	4.46	4.55	5.27	0	2.24	2.51	2.01	3.32
$\mathbf{y}_{12}$	7.42	6.94	7.68	7.60	6.67	4.08	4.42	4.47	4.36	5.42	2.24	0	3.12	2.17	3.49
$\mathbf{y}_{13}$	7.46	7.18	7.27	7.67	7.42	4.08	4.78	3.93	4.54	5.12	2.51	3.12	0	2.23	3.62
$\mathbf{y}_{14}$	7.34	6.80	7.11	7.52	7.03	4.08	4.86	4.25	4.40	5.55	2.01	2.17	2.23	0	3.11
$\mathbf{y}_{15}$	6.71	6.80	6.66	6.97	6.53	4.67	4.29	3.45	3.98	5.08	3.32	3.49	3.62	3.11	0

Table 4.13: The normalized extended Ichino-Yaguchi dissimilarity matrix for variable Y_1.

ϕ_1^*	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.12	0.04	0.05	0.00	0.29	0.27	0.35	0.25	0.20	0.44	0.42	0.46	0.47	0.40
$\mathbf{y}_2$	0.12	0	0.11	0.08	0.11	0.20	0.18	0.26	0.17	0.13	0.32	0.33	0.37	0.39	0.31
$\mathbf{y}_3$	0.04	0.11	0	0.05	0.04	0.27	0.24	0.31	0.22	0.17	0.42	0.40	0.43	0.44	0.37
$\mathbf{y}_4$	0.05	0.08	0.05	0	0.05	0.24	0.21	0.29	0.20	0.14	0.38	0.37	0.41	0.42	0.35
$\mathbf{y}_5$	0.00	0.11	0.04	0.05	0	0.29	0.26	0.34	0.25	0.20	0.43	0.42	0.45	0.47	0.39
$\mathbf{y}_6$	0.29	0.20	0.27	0.24	0.29	0	0.09	0.14	0.13	0.15	0.14	0.19	0.19	0.23	0.15
$\mathbf{y}_7$	0.27	0.18	0.24	0.21	0.26	0.09	0	0.09	0.07	0.09	0.22	0.16	0.18	0.22	0.12
$\mathbf{y}_8$	0.35	0.26	0.31	0.29	0.34	0.14	0.09	0	0.08	0.15	0.20	0.14	0.14	0.17	0.04
$\mathbf{y}_9$	0.25	0.17	0.22	0.20	0.25	0.13	0.07	0.08	0	0.12	0.25	0.19	0.21	0.25	0.14
$\mathbf{y}_{10}$	0.20	0.13	0.17	0.14	0.20	0.15	0.09	0.15	0.12	0	0.29	0.24	0.26	0.29	0.21
$\mathbf{y}_{11}$	0.44	0.32	0.42	0.38	0.43	0.14	0.22	0.20	0.25	0.29	0	0.13	0.08	0.13	0.19
$\mathbf{y}_{12}$	0.42	0.33	0.40	0.37	0.42	0.19	0.16	0.14	0.19	0.24	0.13	0	0.07	0.06	0.14
$\mathbf{y}_{13}$	0.46	0.37	0.43	0.41	0.45	0.19	0.18	0.14	0.21	0.26	0.08	0.07	0	0.06	0.12
$\mathbf{y}_{14}$	0.47	0.39	0.44	0.42	0.47	0.23	0.22	0.17	0.25	0.29	0.13	0.06	0.06	0	0.14
$\mathbf{y}_{15}$	0.40	0.31	0.37	0.35	0.39	0.15	0.12	0.04	0.14	0.21	0.19	0.14	0.12	0.14	0

Table 4.14: The normalized extended Ichino-Yaguchi dissimilarity matrix for variable Y_2.

ϕ_2^*	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.12	0.08	0.13	0.11	0.19	0.27	0.20	0.22	0.24	0.13	0.17	0.16	0.15	0.12
$\mathbf{y}_2$	0.12	0	0.10	0.12	0.16	0.31	0.40	0.33	0.35	0.37	0.28	0.28	0.28	0.23	0.25
$\mathbf{y}_3$	0.08	0.10	0	0.11	0.13	0.21	0.30	0.22	0.25	0.26	0.18	0.21	0.17	0.12	0.15
$\mathbf{y}_4$	0.13	0.12	0.11	0	0.19	0.26	0.34	0.28	0.30	0.31	0.23	0.27	0.23	0.22	0.20
$\mathbf{y}_5$	0.11	0.16	0.13	0.19	0	0.19	0.25	0.22	0.21	0.24	0.16	0.11	0.21	0.15	0.16
$\mathbf{y}_6$	0.19	0.31	0.21	0.26	0.19	0	0.16	0.10	0.04	0.10	0.12	0.12	0.10	0.10	0.09
$\mathbf{y}_7$	0.27	0.40	0.30	0.34	0.25	0.16	0	0.09	0.13	0.11	0.16	0.14	0.17	0.19	0.15
$\mathbf{y}_8$	0.20	0.33	0.22	0.28	0.22	0.10	0.09	0	0.09	0.04	0.07	0.14	0.08	0.11	0.07
$\mathbf{y}_9$	0.22	0.35	0.25	0.30	0.21	0.04	0.13	0.09	0	0.08	0.10	0.12	0.12	0.13	0.10
$\mathbf{y}_{10}$	0.24	0.37	0.26	0.31	0.24	0.10	0.11	0.04	0.08	0	0.10	0.14	0.12	0.15	0.11
$\mathbf{y}_{11}$	0.13	0.28	0.18	0.23	0.16	0.12	0.16	0.07	0.10	0.10	0	0.11	0.11	0.11	0.06
$\mathbf{y}_{12}$	0.17	0.28	0.21	0.27	0.11	0.12	0.14	0.14	0.12	0.14	0.11	0	0.16	0.13	0.13
$\mathbf{y}_{13}$	0.16	0.28	0.17	0.23	0.21	0.10	0.17	0.08	0.12	0.12	0.11	0.16	0	0.09	0.07
$\mathbf{y}_{14}$	0.15	0.23	0.12	0.22	0.15	0.10	0.19	0.11	0.13	0.15	0.11	0.13	0.09	0	0.09
$\mathbf{y}_{15}$	0.12	0.25	0.15	0.20	0.16	0.09	0.15	0.07	0.10	0.11	0.06	0.13	0.07	0.09	0

Table 4.15: The normalized extended Ichino-Yaguchi dissimilarity matrix for variable Y_3.

ϕ_3^*	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.01	0.03	0.02	0.02	0.47	0.44	0.49	0.46	0.43	0.72	0.67	0.66	0.69	0.64
$\mathbf{y}_2$	0.01	0	0.03	0.02	0.03	0.47	0.44	0.49	0.46	0.43	0.72	0.68	0.66	0.69	0.64
$\mathbf{y}_3$	0.03	0.03	0	0.03	0.03	0.45	0.41	0.46	0.44	0.41	0.69	0.65	0.63	0.67	0.62
$\mathbf{y}_4$	0.02	0.02	0.03	0	0.02	0.47	0.44	0.49	0.46	0.43	0.72	0.68	0.66	0.70	0.64
$\mathbf{y}_5$	0.02	0.03	0.03	0.02	0	0.46	0.43	0.48	0.45	0.42	0.71	0.67	0.65	0.68	0.63
$\mathbf{y}_6$	0.47	0.47	0.45	0.47	0.46	0	0.06	0.02	0.07	0.05	0.24	0.22	0.20	0.22	0.17
$\mathbf{y}_7$	0.44	0.44	0.41	0.44	0.43	0.06	0	0.05	0.05	0.05	0.27	0.24	0.23	0.25	0.19
$\mathbf{y}_8$	0.49	0.49	0.46	0.49	0.48	0.02	0.05	0	0.05	0.07	0.22	0.20	0.18	0.20	0.15
$\mathbf{y}_9$	0.46	0.46	0.44	0.46	0.45	0.07	0.05	0.05	0	0.07	0.24	0.22	0.21	0.22	0.17
$\mathbf{y}_{10}$	0.43	0.43	0.41	0.43	0.42	0.05	0.05	0.07	0.07	0	0.28	0.25	0.23	0.25	0.20
$\mathbf{y}_{11}$	0.72	0.72	0.69	0.72	0.71	0.24	0.27	0.22	0.24	0.28	0	0.07	0.05	0.03	0.10
$\mathbf{y}_{12}$	0.67	0.68	0.65	0.68	0.67	0.22	0.24	0.20	0.22	0.25	0.07	0	0.11	0.11	0.10
$\mathbf{y}_{13}$	0.66	0.66	0.63	0.66	0.65	0.20	0.23	0.18	0.21	0.23	0.05	0.11	0	0.09	0.10
$\mathbf{y}_{14}$	0.69	0.69	0.67	0.70	0.68	0.22	0.25	0.20	0.22	0.25	0.03	0.11	0.09	0	0.07
$\mathbf{y}_{15}$	0.64	0.64	0.62	0.64	0.63	0.17	0.19	0.15	0.17	0.20	0.10	0.10	0.10	0.07	0

Table 4.16: The normalized extended Ichino-Yaguchi dissimilarity matrix for variable Y_4.

ϕ_4^*	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.01	0.04	0.00	0.05	0.46	0.42	0.48	0.44	0.40	0.72	0.73	0.68	0.68	0.78
$\mathbf{y}_2$	0.01	0	0.04	0.01	0.05	0.44	0.41	0.46	0.43	0.38	0.71	0.71	0.67	0.67	0.76
$\mathbf{y}_3$	0.04	0.04	0	0.05	0.03	0.40	0.36	0.41	0.38	0.34	0.65	0.65	0.61	0.61	0.70
$\mathbf{y}_4$	0.00	0.01	0.05	0	0.06	0.47	0.43	0.48	0.45	0.41	0.73	0.73	0.69	0.69	0.79
$\mathbf{y}_5$	0.05	0.05	0.03	0.06	0	0.41	0.37	0.42	0.39	0.35	0.66	0.66	0.62	0.62	0.72
$\mathbf{y}_6$	0.46	0.44	0.40	0.47	0.41	0	0.04	0.04	0.02	0.05	0.27	0.26	0.22	0.23	0.31
$\mathbf{y}_7$	0.42	0.41	0.36	0.43	0.37	0.04	0	0.07	0.04	0.04	0.30	0.30	0.25	0.26	0.34
$\mathbf{y}_8$	0.48	0.46	0.41	0.48	0.42	0.04	0.07	0	0.05	0.08	0.25	0.24	0.20	0.20	0.30
$\mathbf{y}_9$	0.44	0.43	0.38	0.45	0.39	0.02	0.04	0.05	0	0.05	0.29	0.27	0.24	0.24	0.33
$\mathbf{y}_{10}$	0.40	0.38	0.34	0.41	0.35	0.05	0.04	0.08	0.05	0	0.33	0.32	0.27	0.29	0.37
$\mathbf{y}_{11}$	0.72	0.71	0.65	0.73	0.66	0.27	0.30	0.25	0.29	0.33	0	0.11	0.09	0.11	0.10
$\mathbf{y}_{12}$	0.73	0.71	0.65	0.73	0.66	0.26	0.30	0.24	0.27	0.32	0.11	0	0.07	0.03	0.08
$\mathbf{y}_{13}$	0.68	0.67	0.61	0.69	0.62	0.22	0.25	0.20	0.24	0.27	0.09	0.07	0	0.07	0.11
$\mathbf{y}_{14}$	0.68	0.67	0.61	0.69	0.62	0.23	0.26	0.20	0.24	0.29	0.11	0.03	0.07	0	0.10
$\mathbf{y}_{15}$	0.78	0.76	0.70	0.79	0.72	0.31	0.34	0.30	0.33	0.37	0.10	0.08	0.11	0.10	0

Table 4.17: The normalized city block distance matrix.

$\mathbf{D}^1_{NM}$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.26	0.20	0.20	0.19	1.41	1.40	1.50	1.37	1.27	2.02	1.99	1.96	2.00	1.94
$\mathbf{y}_2$	0.26	0	0.28	0.22	0.34	1.43	1.42	1.54	1.42	1.32	2.03	2.00	1.97	1.98	1.97
$\mathbf{y}_3$	0.20	0.28	0	0.25	0.23	1.33	1.31	1.41	1.29	1.18	1.94	1.91	1.84	1.84	1.83
$\mathbf{y}_4$	0.20	0.22	0.25	0	0.32	1.44	1.42	1.54	1.41	1.29	2.06	2.06	1.99	2.03	1.98
$\mathbf{y}_5$	0.19	0.34	0.23	0.32	0	1.34	1.31	1.46	1.30	1.21	1.97	1.86	1.93	1.92	1.90
$\mathbf{y}_6$	1.41	1.43	1.33	1.44	1.34	0	0.35	0.30	0.26	0.35	0.77	0.79	0.70	0.78	0.73
$\mathbf{y}_7$	1.40	1.42	1.31	1.42	1.31	0.35	0	0.31	0.29	0.30	0.96	0.84	0.83	0.92	0.81
$\mathbf{y}_8$	1.50	1.54	1.41	1.54	1.46	0.30	0.31	0	0.27	0.34	0.75	0.72	0.60	0.68	0.56
$\mathbf{y}_9$	1.37	1.42	1.29	1.41	1.30	0.26	0.29	0.27	0	0.32	0.88	0.81	0.78	0.84	0.73
$\mathbf{y}_{10}$	1.27	1.32	1.18	1.29	1.21	0.35	0.30	0.34	0.32	0	0.99	0.95	0.89	0.99	0.90
$\mathbf{y}_{11}$	2.02	2.03	1.94	2.06	1.97	0.77	0.96	0.75	0.88	0.99	0	0.42	0.33	0.38	0.45
$\mathbf{y}_{12}$	1.99	2.00	1.91	2.06	1.86	0.79	0.84	0.72	0.81	0.95	0.42	0	0.42	0.33	0.45
$\mathbf{y}_{13}$	1.96	1.97	1.84	1.99	1.93	0.70	0.83	0.60	0.78	0.89	0.33	0.42	0	0.30	0.40
$\mathbf{y}_{14}$	2.00	1.98	1.84	2.03	1.92	0.78	0.92	0.68	0.84	0.99	0.38	0.33	0.30	0	0.41
$\mathbf{y}_{15}$	1.94	1.97	1.83	1.98	1.90	0.73	0.81	0.56	0.73	0.90	0.45	0.45	0.40	0.41	0

Table 4.18: The normalized Euclidean distance matrix.

$\mathbf{D}^2_{NM}$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.17	0.11	0.14	0.12	0.74	0.72	0.79	0.72	0.67	1.12	1.09	1.07	1.09	1.09
$\mathbf{y}_2$	0.17	0	0.16	0.14	0.20	0.75	0.74	0.79	0.74	0.70	1.10	1.07	1.05	1.06	1.08
$\mathbf{y}_3$	0.11	0.16	0	0.14	0.14	0.69	0.67	0.73	0.67	0.62	1.05	1.02	0.99	1.01	1.01
$\mathbf{y}_4$	0.14	0.14	0.14	0	0.21	0.75	0.73	0.80	0.74	0.69	1.12	1.10	1.06	1.09	1.09
$\mathbf{y}_5$	0.12	0.20	0.14	0.21	0	0.70	0.67	0.75	0.68	0.63	1.07	1.03	1.03	1.05	1.04
$\mathbf{y}_6$	0.74	0.75	0.69	0.75	0.70	0	0.20	0.18	0.15	0.19	0.41	0.41	0.36	0.41	0.40
$\mathbf{y}_7$	0.72	0.74	0.67	0.73	0.67	0.20	0	0.16	0.16	0.16	0.49	0.44	0.42	0.46	0.44
$\mathbf{y}_8$	0.79	0.79	0.73	0.80	0.75	0.18	0.16	0	0.14	0.19	0.40	0.37	0.31	0.35	0.34
$\mathbf{y}_9$	0.72	0.74	0.67	0.74	0.68	0.15	0.16	0.14	0	0.17	0.46	0.42	0.40	0.43	0.40
$\mathbf{y}_{10}$	0.67	0.70	0.62	0.69	0.63	0.19	0.16	0.19	0.17	0	0.53	0.49	0.46	0.51	0.49
$\mathbf{y}_{11}$	1.12	1.10	1.05	1.12	1.07	0.41	0.49	0.40	0.46	0.53	0	0.21	0.17	0.21	0.24
$\mathbf{y}_{12}$	1.09	1.07	1.02	1.10	1.03	0.41	0.44	0.37	0.42	0.49	0.21	0	0.22	0.18	0.23
$\mathbf{y}_{13}$	1.07	1.05	0.99	1.06	1.03	0.36	0.42	0.31	0.40	0.46	0.17	0.22	0	0.15	0.20
$\mathbf{y}_{14}$	1.09	1.06	1.01	1.09	1.05	0.41	0.46	0.35	0.43	0.51	0.21	0.18	0.15	0	0.21
$\mathbf{y}_{15}$	1.09	1.08	1.01	1.09	1.04	0.40	0.44	0.34	0.40	0.49	0.24	0.23	0.20	0.21	0

Table 4.19: The normalized CDF dissimilarity matrix.

$\mathbf{D}_{NCDF}$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_6$	$\mathbf{y}_7$	$\mathbf{y}_8$	$\mathbf{y}_9$	$\mathbf{y}_{10}$	$\mathbf{y}_{11}$	$\mathbf{y}_{12}$	$\mathbf{y}_{13}$	$\mathbf{y}_{14}$	$\mathbf{y}_{15}$
$\mathbf{y}_1$	0	0.27	0.34	0.13	0.21	1.42	1.40	1.52	1.39	1.28	2.04	1.99	1.95	1.98	1.95
$\mathbf{y}_2$	0.27	0	0.43	0.17	0.38	1.45	1.43	1.55	1.42	1.31	2.07	2.02	1.98	2.01	1.98
$\mathbf{y}_3$	0.34	0.43	0	0.35	0.23	1.35	1.34	1.42	1.32	1.24	1.81	1.80	1.78	1.78	1.74
$\mathbf{y}_4$	0.13	0.17	0.35	0	0.28	1.43	1.42	1.54	1.41	1.29	2.05	2.01	1.97	2.00	1.96
$\mathbf{y}_5$	0.21	0.38	0.23	0.28	0	1.36	1.35	1.46	1.33	1.23	1.91	1.86	1.88	1.88	1.83
$\mathbf{y}_6$	1.42	1.45	1.35	1.43	1.36	0	0.24	0.17	0.22	0.29	0.70	0.63	0.63	0.69	0.66
$\mathbf{y}_7$	1.40	1.43	1.34	1.42	1.35	0.24	0	0.25	0.13	0.18	0.89	0.77	0.79	0.86	0.83
$\mathbf{y}_8$	1.52	1.55	1.42	1.54	1.46	0.17	0.25	0	0.23	0.32	0.67	0.59	0.55	0.62	0.59
$\mathbf{y}_9$	1.39	1.42	1.32	1.41	1.33	0.22	0.13	0.23	0	0.18	0.85	0.73	0.74	0.81	0.78
$\mathbf{y}_{10}$	1.28	1.31	1.24	1.29	1.23	0.29	0.18	0.32	0.18	0	0.95	0.87	0.87	0.94	0.91
$\mathbf{y}_{11}$	2.04	2.07	1.81	2.05	1.91	0.70	0.89	0.67	0.85	0.95	0	0.25	0.22	0.18	0.31
$\mathbf{y}_{12}$	1.99	2.02	1.80	2.01	1.86	0.63	0.77	0.59	0.73	0.87	0.25	0	0.28	0.22	0.29
$\mathbf{y}_{13}$	1.95	1.98	1.78	1.97	1.88	0.63	0.79	0.55	0.74	0.87	0.22	0.28	0	0.16	0.28
$\mathbf{y}_{14}$	1.98	2.01	1.78	2.00	1.88	0.69	0.86	0.62	0.81	0.94	0.18	0.22	0.16	0	0.29
$\mathbf{y}_{15}$	1.95	1.98	1.74	1.96	1.83	0.66	0.83	0.59	0.78	0.91	0.31	0.29	0.28	0.29	0

Diday measure, the normalized city block distance, the normalized Euclidean distance, and the normalized CDF distance and the monothetic and polythetic clustering methods are shown in Figure 4.4, 4.5, 4.6, and 4.7, respectively.

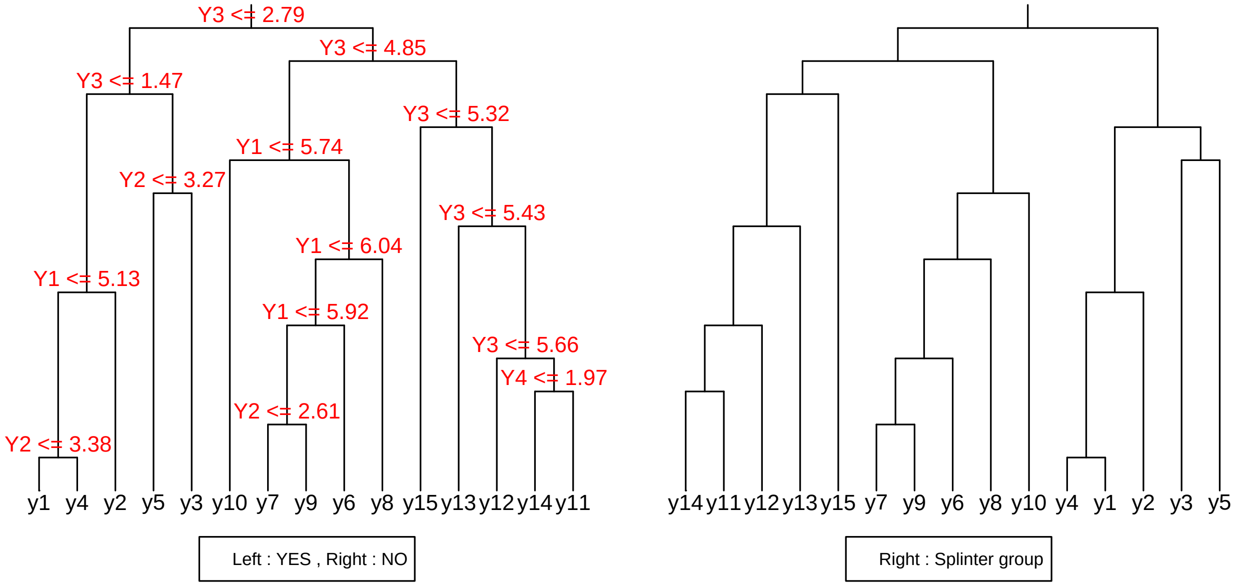
Figure 4.4 shows the clustering results for the extended Gowda-Diday measure, for the monothetic and polythetic clustering methods. From Figure 4.4(a), we see that for the monothetic method there are 14 binary questions, and the binary question at the first stage is 'Is $Y_3 \leq 2.79?'$. *This means that bipartiotioning is based on variable* $Y_3 =$*'Petal Length', and the cut point for 'Yes' or 'No' is 2.79. Thus, if the answer is 'Yes', the iris species goes to the cluster* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$. *Conversely, if 'No', it goes to* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$. *That is, if the Petal Length is equal or less than 2.79, then the iris species goes to the setosa, and if the Petal Length is larger than 2.79, then it goes to the versicolor and virginica. For example, suppose that there is a histogram-valued object for a new iris species and the mean for the variable* Y_3 *of this object,* $M_{ij} = M_{i3} = 2.2$ *of Equation (3.13). Then, since the mean of this new iris species object is less than the cut point (2.79), the answer is 'Yes' and this new object is classified into the cluster* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$ *(setosa). At the second stage, the binary question is 'Is* $Y_3 \leq 4.85?'$. *Similarly to the first stage, if the answer is 'Yes', then it goes to the cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}$, *and if 'No', then it goes to* $\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}$ *(virginica). For example, suppose that we have a histogram-valued object for a new iris species, and the mean for the variable* Y_3 *of a new histogram-valued object is 4.0. Then, the answer for the first binary question, 'Is* $Y_3 \leq 2.79?'$, *is 'No', and the answer for the second binary question, 'Is* $Y_3 \leq 4.85?'$, *is 'Yes'. Thus, this object is classified into the cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}$ *(versicolor).*

As shown in Figure 4.4(b), the polythetic method does not provide binary questions because it uses all p variables to find the optimal bipartition. From Figure 4.4(b), the right side of each node represents the splinter cluster. The polythetic method proposed in Section 3.4.2 starts from finding the object that is the most different from the others within a cluster. That object is called the seed, and the cluster including the seed is called the splinter cluster or group. The polythetic method iteratively compares whether each object is close to a main *cluster or a splinter cluster. Thus, from Figure 4.4(b), we know that the splinter cluster for the first stage is* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$*. At the second stage, the cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *is bipartitioned into* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}$ *and* $\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}$*. In this case,* $\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}$ *is the main cluster and* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}$ *is the splinter cluster.*

Figure 4.5 shows the dendrograms obtained when using the normalized city block distance with $q = 1$ *in Equation (3.28). Figure 4.5(a) is obtained using the monothetic clustering method and Figure 4.5(b) uses the polythetic method. The monothetic methods in Figure 4.5(b) provides the binary questions. At the first stage,* $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_{15}\}$ *is bipartitioned into* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$ *and* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$*. The first binary question is 'Is* $Y_3 \leq 2.79$*?'. This means that if the Petal Length is less than 2.78, then it belongs to the cluster* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$ *(setosa), and otherwise, it goes to* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *(versicolor and virginica). The binary question for the second stage is 'Is* $Y_4 \leq 1.63$*?'. The cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *is bipartitioned into* $(\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}$ *(versicolor) and* $\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}$ *(virginica) by the variable* $Y_4 =$*'Petal Width'. The bipartitions for the first and second stage are the exactly same as those for the extended Gowda-Diday measure as shown in Figure 4.4(a), but the binary questions for the second stage are different.*

From Figure 4.5(b), we know that the splinter cluster for the first stage is $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$*. At the second stage, the cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *is bipartitioned into* $\big(\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\},\ \{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}\big)$*, and the splinter cluster is* $\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}$*. The bipartitions for the first and second stage are the same as those of the monothetic method of Figure 4.4(a).*

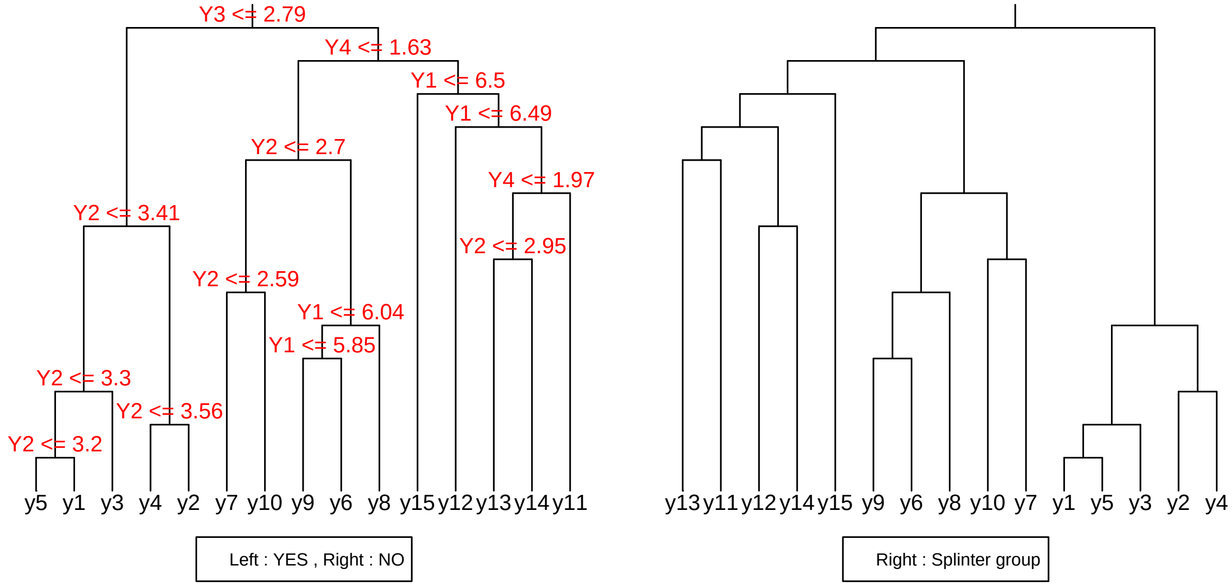
Figure 4.6 shows the dendrograms obtained when using the normalized Euclidean distance with $q = 2$ *in Equation (3.28). Figure 4.6(a) is obtained using the monothetic clustering algorithm and Figure 4.6(b) comes from the polythetic algorithm. From Figure 4.6(a) by the monothetic method, the first binary question is 'Is* $Y_3 \leq 2.79$*?', and* $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_{15}\}$ *is bipartitioned into* $\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}$ *and* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *by this binary question. At the second stage, the cluster* $\{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}$ *is split into* $\big(\{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\},\ \{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}\big)$ *by the binary question 'Is* $Y_4 \leq 1.63$*?'. For example, suppose that there is a new histogram-valued object, and the mean for* $Y_3 =$*'Petal Length' of this object is 3.5 and the mean for* $Y_4 =$*'Petal Width' is 2.0.*



(a) Monothetic method

(b) Polythetic method

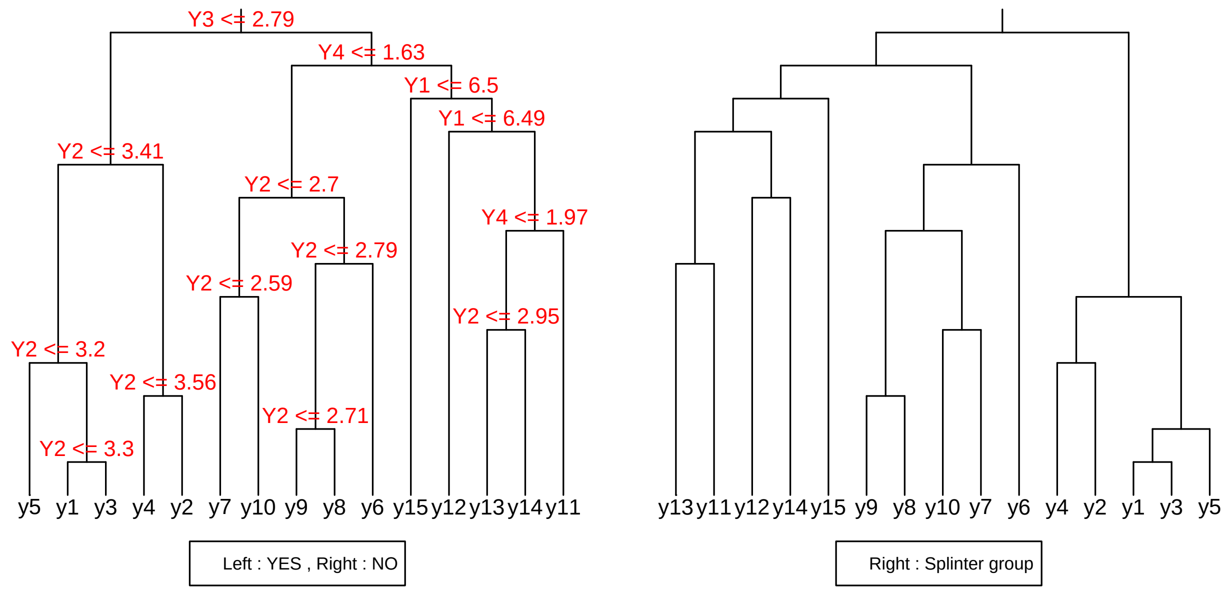
Figure 4.4: Dendrogram for the extended Gowda-Diday measure.



(a) Monothetic method

(b) Polythetic method

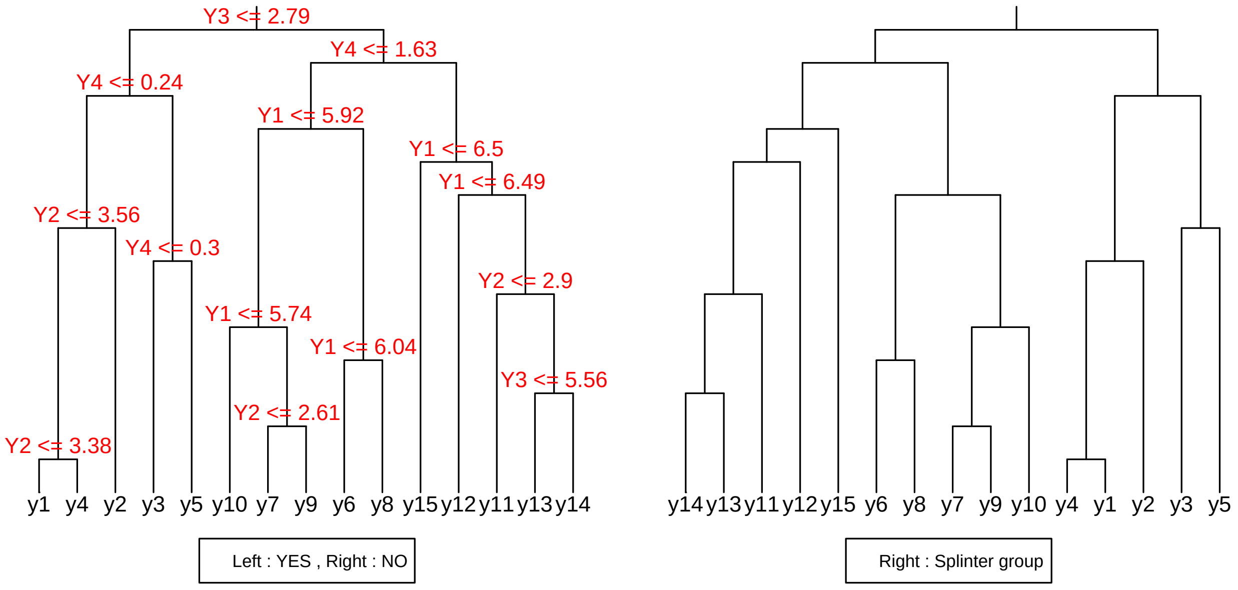
Figure 4.5: Dendrogram for the normalized city block measure.



(a) Monothetic method

(b) Polythetic method

Figure 4.6: Dendrogram for the normalized Euclidean measure.



(a) Monothetic method

(b) Polythetic method

Figure 4.7: Dendrogram for the normalized CDF measure.

Then, this object goes to the cluster $\{\mathbf{y}_{11},\ldots,\mathbf{y}_{15}\}$ *(virginica). The partitions for the first and second stage are the exactly same as those obtained for the extended Gowda-Diday measure and for the normalized city block distance.*

From Figure 4.6(b), we know that $\Omega = \{\mathbf{y}_1,\ldots,\mathbf{y}_{15}\}$ *is bipartitioned into* $\{\mathbf{y}_1,\ldots,\mathbf{y}_5\}$ *and* $\{\mathbf{y}_6,\ldots,\mathbf{y}_{15}\}$ *at the first stage, and the splinter cluster is* $\{\mathbf{y}_1,\ldots,\mathbf{y}_5\}$*. At the second stage, the cluster* $\{\mathbf{y}_6,\ldots,\mathbf{y}_{15}\}$ *is split into* $\big(\{\mathbf{y}_6,\ldots,\mathbf{y}_{10}\},\ \{\mathbf{y}_{11},\ldots,\mathbf{y}_{15}\}\big)$*. These bipartitions for the first and second stage are the same as those obtained by the monothetic method shown in Figure 4.6(a).*

Figure 4.7 shows the dendrograms obtained when using the normalized CDF dissimilarity measure of Equation (3.31). Figure 4.7(a) is obtained using the monothetic clustering method and Figure 4.7(b) uses the polythetic method. From Figure 4.7(a) for the monothetic method, $\Omega = \{\mathbf{y}_1,\ldots,\mathbf{y}_{15}\}$ *is bipartitioned into* $\{\mathbf{y}_1,\ldots,\mathbf{y}_5\}$ *and* $\{\mathbf{y}_6,\ldots,\mathbf{y}_{15}\}$ *at the first stage, the binary question for this stage is 'Is* $Y_3 \leq 2.79$*?'. At the second stage, the cluster* $\{\mathbf{y}_6,\ldots,\mathbf{y}_{15}\}$ *is split into* $\big(\{\mathbf{y}_6,\ldots,\mathbf{y}_{10}\},\ \{\mathbf{y}_{11},\ldots,\mathbf{y}_{15}\}\big)$ *by the binary question 'Is* $Y_4 \leq 1.63$*?'. These bipartitions for the first and second stage are the same as those of the extended Gowda-Diday, the normalized city block distance, and the normalized Euclidean distance. From Figure 4.7(b) by the polythetic method, we also know that the partitions for the first and second stage are same as those of the monothetic method shown in Figure 4.7(a).*

Thus, when the numbers of clusters are two and three, we know that all dissimilarity measures and clustering methods give the same results. That is, when $r = 2$*, the clusters are* $\big(\{\mathbf{y}_1,\ldots,\mathbf{y}_5\},\{\mathbf{y}_6,\ldots,\mathbf{y}_{15}\}\big)$*; and the clusters are* $\big(\{\mathbf{y}_1,\ldots,\mathbf{y}_5\},\{\mathbf{y}_6,\ldots,\mathbf{y}_{10}\},\{\mathbf{y}_{11},\ldots,\mathbf{y}_{15}\}\big)$ *when* $r = 3$*. This partition coincides with the attribute of the original classical iris dataset. From the distance or dissimilarity matrices in Table 4.12, 4.17, 4.18, and 4.19 and the clustering outcomes in Figure 4.4, 4.5, 4.6, and 4.7, we can calculate the Dunn and Davis-Bouldin indexes for symbolic objects using Equation (4.8) and (4.9).*

Table 4.20 shows the cluster validity index values for the extended Gowda-Diday measure, the normalized city block distance, the normalized Euclidean distance, and the normalized

Table 4.20: Cluster validity index values for Example 4.3

Measure	Method	Index	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 6$	$r = 7$	$r = 8$	$r = 9$	$r = 10$	$r = 11$	$r = 12$	$r = 13$	$r = 14$
Extended Gowda-Diday	Monothetic	DI_r^s	1.797	*2.216*	0.476	0.589	0.568	0.537	0.533	0.704	0.587	0.578	1.265	1.124	1.008
		DB_r^s	0.698	*0.688*	1.733	2.125	1.985	2.064	1.936	1.457	1.562	1.494	0.814	0.735	0.714
	Polythetic	DI_r^s	1.797	*2.216*	0.442	0.589	0.370	0.537	0.533	0.704	0.587	0.575	1.265	1.124	1.008
		DB_r^s	0.698	*0.688*	2.181	2.125	2.416	2.064	1.936	1.457	1.562	1.542	0.814	0.735	0.714
Normalized city block	Monothetic	DI_r^s	*4.610*	4.595	0.521	0.450	0.552	0.541	0.812	0.597	0.593	0.998	0.758	1.040	1.426
		DB_r^s	*0.232*	0.268	1.667	1.709	1.842	1.817	1.438	1.560	1.616	1.006	1.010	0.723	0.482
	Polythetic	DI_r^s	*4.610*	4.595	0.521	0.487	0.283	0.420	0.349	0.347	0.333	0.998	0.758	0.574	1.426
		DB_r^s	*0.232*	0.268	1.667	1.595	2.283	2.045	2.191	2.412	2.479	1.006	1.010	1.113	0.482
Normalized Euclidean	Monothetic	DI_r^s	*4.804*	4.573	0.493	0.417	0.393	0.662	0.814	0.928	0.797	0.740	1.012	1.028	1.640
		DB_r^s	*0.227*	0.282	1.739	1.841	2.530	1.930	1.508	1.153	1.136	1.208	0.915	0.857	0.505
	Polythetic	DI_r^s	*4.804*	4.573	0.493	0.432	0.379	0.334	0.304	0.242	0.740	0.740	0.623	0.606	1.640
		DB_r^s	*0.227*	0.282	1.739	1.778	2.421	2.782	2.847	3.097	1.238	1.208	1.315	1.302	0.505
Normalized CDF	Monothetic	DI_r^s	5.556	*7.909*	0.822	0.494	0.729	0.980	0.881	0.768	0.805	1.359	1.077	1.434	1.031
		DB_r^s	0.199	*0.163*	0.852	1.617	1.419	1.023	1.070	1.144	1.004	0.666	0.669	0.520	0.530
	Polythetic	DI_r^s	5.556	*7.909*	0.822	0.515	0.371	0.980	0.681	0.768	0.805	1.359	1.077	1.434	1.031
		DB_r^s	0.199	*0.163*	0.852	1.653	1.967	1.023	1.261	1.144	1.004	0.666	0.669	0.520	0.530

CDF measure, and the monothetic and polythetic divisive clustering methods. The monothetic and polythetic methods give the same maximum value of the Dunn index and the same minimum value of the Davis-Bouldin index. The validity indexes for the extended Gowda-Diday and normalized CDF measures show that the optimal number of clusters is three because the Dunn index value for these measures is largest at $r = 3$ *and the Davis-Bouldin index value for* $r = 3$ *is smallest. On the contrary, the index values for the normalized city block distance and the normalized Euclidean distance indicate that there are two clusters in this dataset. However, all index values at* $r = 2$ *are relatively very close to the index values at* $r = 3$*. If there are two clusters, the clusters are* $\big(\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}, \{\mathbf{y}_6, \ldots, \mathbf{y}_{15}\}\big)$*, and if there are three clusters, we have* $\big(\{\mathbf{y}_1, \ldots, \mathbf{y}_5\}, \{\mathbf{y}_6, \ldots, \mathbf{y}_{10}\}, \{\mathbf{y}_{11}, \ldots, \mathbf{y}_{15}\}\big)$*. This result coincides with the iris species in the original dataset. This example shows that the monothetic and polythetic divisive clustering methods, and the extended Gowda-Diday measure, the normalized city block distance, the normalized Euclidean distance, and the normalized CDF measure give a good result for clustering in this dataset, and the validity indexes find the optimal number of clusters well.*

From these examples, we know that the Dunn index, DI_r^s, and the Davais-Bouldin index, DB_r^s, proposed in this chapter, work well for symbolic objects. They can be useful indexes that give the information for the optimal number of clusters and help to choose well-separated partitions.

Chapter 5

Data Analysis

In this chapter, the effectiveness of the divisive clustering methods proposed in this study is demonstrated on the forest cover type dataset with cartographic variables (available in the UCI Machine Learning Repository web site; http://www.ics.uci.edu/∼mlearn/ MLRepository.html). This dataset includes 581,012 individual observations and ten numeric variables with information for four wilderness areas, viz., the Rawah, Comanche Peak, Neota, and Cache la Poudre in the Roosevelt National Forest, located in northern Colorado. The forest cover type dataset came from the US Forest Service inventory information. The cartographic variables in the dataset are the location information for seven cover types such as elevation, aspect, and slope. Most variables were derived from standard digital spatial data precessed in a geographical information system (GIS).

Originally, this dataset was made up for a discriminant analysis of cover types using the geographical and environmental information. However, the goal of the analysis in this study is to investigate which forest cover types have the most similar location and environment to each other. Since this dataset has a huge number of individual observations and we are interested in cover types, in order to obtain the answer for the research question it is appropriate that data are aggregated by each cover type. This aggregation transforms the original data into histogram-valued data. That is, each cover type is regarded as a histogram-valued object. A more detailed description of the data is given in Section 5.1. Thus, we perform clustering for histogram-valued data using various dissimilarity measures. Through the analysis for the forest cover type data, we show in Section 5.2 and 5.3 how to apply the dissimilarity measures and the divisive clustering algorithms proposed in this study to real data sets. Then, in Section 5.4, we examine whether they work well or not.

5.1 Data Description

The cover type dataset consists of information for four wilderness areas in the Roosevelt National Forest, Colorado. The dataset has 581,012 individual observations and ten numeric variables including information for the place and environment where each cover type is located. Each observation has one of seven mutually exclusive forest cover type classes. Table 5.1 shows the cover type classes and the number of individual observations for each class. For example, the Spruce-fir cover type has 211,840 observations. The lodgepole pine and Spruce-fir have the largest number of observations relatively to the other cover types, and the cottonwood/willow cover type has the smallest number of observations (at 2,747). For the purposes of our analysis, we will not consider original frequencies for each object in the symbolic data context. Thus, each cover type is treated as one observation. An extension of our methodology could weight each observation, e.g., in proportion to these frequencies.

Table 5.1: Forest cover type classes.

Symbolic object	Forest cover type	# of obs.
$\mathbf{y}_1$	Spruce-fir	211,840
$\mathbf{y}_2$	Lodgepole pine	283,301
$\mathbf{y}_3$	Ponderosa pine	35,754
$\mathbf{y}_4$	Cottonwood/Willow	2,747
$\mathbf{y}_5$	Aspen	9,493
$\mathbf{y}_6$	Douglas-fir	17,367
$\mathbf{y}_7$	Krummholz	20,510
	Total	581,012

The ten numeric variables in this dataset are digital spatial data obtained from the US Geological Survey (USGS) and the US Forest Service (USFS). The description of each variable is shown in Table 5.2; an expanded description is given in the UCI Machine Learning Repository web site (http://archive.ics.uci.edu/ml/datasets/Covertype). Table 5.3 shows some basic descriptive statistics for each variable of the classical cover type data, viz., the sample mean, standard deviation, and the minimum and maximum values. The measurement scale of five of the variables is in meters, one is in degrees, another is azimuth (i.e., the horizontal angular distance from a fixed reference point), and three are index variables. Since the measurement units of the variables in the dataset are not the same and the ranges of the variables are different, normalized dissimilarity measures are used for clustering.

Table 5.2: Description of each variable in the cover type dataset.

Variable	Measurement	Description
Y_1	Meters	Elevation
Y_2	Azimuth	Aspect
Y_3	Degrees	Slope
Y_4	Meters	Horizontal distance to nearest surface water feature
Y_5	Meters	Vertical distance to nearest surface water feature
Y_6	Meters	Horizontal distance to nearest roadway
Y_7	0 to 255 index	A relative measure of incident sunlight at 09:00 A.M. on the summer solstice
Y_8	0 to 255 index	A relative measure of incident sunlight at noon on the summer solstice
Y_9	0 to 255 index	A relative measure of incident sunlight at 03:00 P.M. on the summer solstice
Y_{10}	Meters	Horizontal distance to nearest historic wildfire ignition point

This dataset is transformed into a histogram-valued dataset with seven symbolic objects and ten histogram-valued variables. To transform the classical cover type data into histogram-valued dataset, we firstly aggregate classical data by each cover type. Then, we have seven groups of aggregated data because there are seven cover types. From each group of aggregated data, we generate one-dimensional histograms for each variable using the ‘hist()’ function in statistical software ‘R’. The number of bins for each histogram is decided by using a formula proposed by Sturges (1926). Sturges’ (1926) formula is $\lceil \log_2 N + 1 \rceil$, where N is the number of classical observations, and $\lceil \cdot \rceil$ is the ceiling function mapping a real number to the next largest integer. This histogram-valued dataset generated from original classical cover type data is provided in Table B.1 of Appendix B.1. We use the resulting histogram-valued dataset for our clustering methodology.

In Section 5.2 and 5.3, we analyse the forest cover type data by performing the clustering methods for symbolic objects proposed in this study. As shown in Table B.1, the histogram-valued data that come from the original cover type data set have different numbers and lengths of subintervals for each object. Thus, in order to obtain the same number and length of subintervals for each object, the histogram-valued data should be transformed using the method introduced in Section 3.2. For each variable, we obtain the starting and ending points of transformed subintervals using Equation (3.3) and (3.4). From the interval with these starting and ending points, the transformed subintervals can be obtained by dividing

Table 5.3: Descriptive statistics for each variable.

Variable	Minimum	Maximum	Mean	Standard Deviation
Y_1	1859	3858	2959.36	279.98
Y_2	0	360	155.65	111.91
Y_3	0	66	14.10	7.49
Y_4	0	1397	269.43	212.55
Y_5	-173	601	46.42	58.30
Y_6	0	7117	2350.15	1559.25
Y_7	0	254	212.15	26.77
Y_8	0	254	223.32	19.77
Y_9	0	254	142.53	38.27
Y_{10}	0	7173	1980.29	1324.19

this interval by the minimum length of original subintervals as defined in Equation (3.5) and (3.7). For each object, the transformed relative frequencies corresponding to each transformed subinterval are determined by the overlapping portion between the original and transformed subintervals. Thus, we can obtain the transformed histogram-valued data from the original histogram-valued data. These transformed data are shown in Table B.2 of Appendix B.2.

5.2 Dissimilarity Measures

From the transformed histogram-valued data of Table B.2, dissimilarity measure values for pairs of objects can be calculated. For the analysis, we apply four different dissimilarity measures to the transformed dataset, viz., the extended Gowda-Diday dissimilarity measure (GD) of Equation (3.21), the city block distance (CB) of Equation (3.26) and the Euclidean distances (EU) of Equation (3.27) (based on the normalized extended Ichino-Yaguchi measure for $\gamma = 0.25$ defined in Equation (3.22) and (3.23)), and the normalized CDF measure (NCDF) of Equation (3.31). All four measures are normalized measures. Details of their derivations are given in Appendix B.3; the corresponding complete dissimilarity/distance matrices $\mathbf{D}_{GD}$, $\mathbf{D}^1_{NM}$, $\mathbf{D}^2_{NM}$, and $\mathbf{D}_{NCDF}$ are as follows:

The extended Gowda-Diday dissimilarity matrix for seven objects of cover types is given by

$$\mathbf{D}_{GD} = \begin{pmatrix} 0.000 & 1.341 & 6.454 & 7.305 & 4.300 & 6.354 & 2.953 \\ & 0.000 & 6.201 & 7.239 & 3.891 & 6.285 & 3.593 \\ & & 0.000 & 4.037 & 4.775 & 2.077 & 7.042 \\ & & & 0.000 & 4.818 & 4.416 & 7.251 \\ & & & & 0.000 & 5.175 & 4.770 \\ & & & & & 0.000 & 7.288 \\ & & & & & & 0.000 \end{pmatrix}. \tag{5.1}$$

The normalized city block and Euclidean distances are based on the normalized extended Ichino-Yaguchi measure (see Appendix B.3), and are special cases of the normalized Minkowski distance of Equation (3.28). If $q = 1$ in the normalized Minkowski distance, then it becomes the normalized city block distance, and if $q = 2$, it becomes the normalized Euclidean distance. The normalized city block distance matrix for seven objects of cover types is

$$\mathbf{D}^1_{NM} = \begin{pmatrix} 0.000 & 0.257 & 1.155 & 1.365 & 0.732 & 1.122 & 0.472 \\ & 0.000 & 1.087 & 1.262 & 0.615 & 1.095 & 0.626 \\ & & 0.000 & 0.589 & 0.743 & 0.348 & 1.323 \\ & & & 0.000 & 0.732 & 0.695 & 1.475 \\ & & & & 0.000 & 0.846 & 0.925 \\ & & & & & 0.000 & 1.364 \\ & & & & & & 0.000 \end{pmatrix}. \tag{5.2}$$

The normalized Euclidean distance matrix for seven objects of cover types is given by

$$\mathbf{D}^2_{NM} = \begin{pmatrix} 0.000 & 0.114 & 0.471 & 0.567 & 0.273 & 0.453 & 0.174 \\ & 0.000 & 0.417 & 0.501 & 0.223 & 0.406 & 0.263 \\ & & 0.000 & 0.212 & 0.279 & 0.133 & 0.568 \\ & & & 0.000 & 0.326 & 0.276 & 0.664 \\ & & & & 0.000 & 0.315 & 0.369 \\ & & & & & 0.000 & 0.565 \\ & & & & & & 0.000 \end{pmatrix}. \tag{5.3}$$

The normalized CDF dissimilarity matrix is

$$\mathbf{D}_{NCDF} = \begin{pmatrix} 0.000 & 0.225 & 1.112 & 1.354 & 0.727 & 1.107 & 0.380 \\ & 0.000 & 1.016 & 1.221 & 0.590 & 1.042 & 0.492 \\ & & 0.000 & 0.596 & 0.702 & 0.330 & 1.291 \\ & & & 0.000 & 0.674 & 0.714 & 1.478 \\ & & & & 0.000 & 0.806 & 0.868 \\ & & & & & 0.000 & 1.364 \\ & & & & & & 0.000 \end{pmatrix}. \tag{5.4}$$

5.3 Clustering

Both the monothetic and polythetic methods are used for divisive clustering, and clustering outcomes are demonstrated using validity indexes (in Section 5.4). These clustering procedures for this cover type dataset are described in Section 5.3.1 in detail; and a discussion of the resulting partitions and dendrograms is in Section 5.3.2.

5.3.1 Monothetic and Polythetic Methods

In this section, we illustrate the monothetic and polythetic clustering procedures for the cover type dataset using the extended Gowda-Diday dissimilarity matrix of Equation (5.1).

Monothetic Method

We illustrate the clustering procedure for the cover type data by the monothetic method. At the first stage, we have a cluster with seven histogram-valued objects for the cover type data. That is,

$$P_1 \equiv C_1 = \{\mathbf{y}_1, \dots, \mathbf{y}_7\},$$

where $\mathbf{y}_i$ are as described in Table 5.1. First, we calculate the mean of each object for each variable, M_{ij}, using Equation (3.13), and then sort the objects in ascending order by their mean values. The result is shown in Table B.3 in Appendix B.4. Let $\mathbf{y}^j_{(i)}$ denote an object with the i^{th} smallest mean for the variable Y_j. Then, from this table, the object with the smallest mean for the variable Y_1 = Elevation is $\mathbf{y}_4$, and this object $\mathbf{y}_4$ is denoted by $\mathbf{y}^1_{(1)}$. The second smallest mean for the variable Y_1 is 2394.2 corresponding to $\mathbf{y}_{(2)} \equiv \mathbf{y}_3$. Also, the object with the largest mean for Y_1 is $\mathbf{y}_7$ and its mean is 3361.7. Thus, Table B.3 shows the order of objects sorted by mean values for each variable.

In the monothetic context, the number of possible bipartitions for the r^{th} stage is $\sum_{u=1}^{r} p(n_u - 1)$, where p is the number of variables and n_u is the number of objects in cluster $C_u,\ u = 1, \dots, r$. Thus, since we have seven objects in C_1, there are 60 $(= 10(7-1))$ possible bipartitions at the first stage. We have to examine the within-cluster variance for all 60 possible bipartitions. For example, from the first row of Table B.3 in Appendix B.4, the possible bipartitions for variable Y_1 are $(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$, $(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$, $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$, $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$, $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_7\})$, and $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_7\})$. For all 60 bipartitions, the decrement values of the within-cluster variance are obtained by Equation(3.39), and the optimal bipartition can be found by the maximum decrement value. For example, if the extended Gowda-Diday dissimilarity matrix and weights $w_i = 1/n$ for the within-cluster variance are used, the decrement value of the within-cluster variance for the partition $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$ can be obtained as follows: Let $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}$, $C_1^1 = \{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}$, and $C_1^2 = \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\}$. Then,

$$\begin{aligned}
I(C_1) &= \frac{1}{7\times 7}\Big\{1.341^2 + 6.454^2 + 7.305^2 + \cdots + 5.175^2 + 4.770^2 + 7.288^2\Big\} \\
&= 12.53, \\
I(C_1^1) &= \frac{1}{7\times 3}\Big\{4.037^2 + 4.416^2 + 2.077^2\Big\} = 1.91, \\
I(C_1^2) &= \frac{1}{7\times 4}\Big\{3.891^2 + 4.300^2 + 4.770^2 + 1.341^2 + 3.593^2 + 2.953^2\Big\} \\
&= 2.85.
\end{aligned}$$

Thus,

$$\Delta_1 = I(C_1) - I(C_1^1) - I(C_1^2) = 12.53 - 1.91 - 2.85 = 7.77.$$

Similarly, we can obtain decrement values of the other possible bipartitions for the first stage, and this result is shown in Table B.4 of Appendix B.4. From the Table B.4, we know that the maximum decrement value is 7.77, and the optimal bipartition corresponding to this value is $(\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\})$.

The important characteristic is that a binary question can be found at each stage. In the case of histogram-valued data, cut points for binary questions can be obtained by the mean value of the union between two objects on a boundary of the optimal bipartition. However, at each stage, there can often exist more than one cut point. For example, as shown in

Table B.4, the clustering outcomes based on the variables Y_1, Y_3, Y_4, Y_6, Y_8, Y_{10}, respectively, detect the same optimal bipartition, $(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$, corresponding to the largest decrement value (7.77). Thus, there can be six cut points at the first stage because six variables detect the optimal bipartition. To obtain a unique cut point and binary question, we use dissimilarity values for these variables between two objects on the boundary of the optimal bipartition. As shown in Table B.4, the optimal bipartition for the variable Y_1 is $(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$. In this case, two objects on the boundary of the optimal bipartition in ascending order by Y_1 are $\mathbf{y}_6$ and $\mathbf{y}_5$. That is, the cluster $\{\mathbf{y}_1, \ldots, \mathbf{y}_7\}$ is split into two clusters in the boundary between two objects $\mathbf{y}_6$ and $\mathbf{y}_5$. Thus, the dissimilarity value for variable Y_1 between two transformed objects $\mathbf{y}'_6$ and $\mathbf{y}'_5$ is considered. That is, from Definition 3.8,

$$\begin{aligned} D_{GD}(y'_{61}, y'_{51}) &= [D_{11}(y'_{61}, y'_{51}) + D_{21}(y'_{61}, y'_{51}) + D_{31}(y'_{61}, y'_{51})] \\ &= 0.319 + 0.725 + 0.180 = 1.223. \end{aligned}$$

Similarly, the dissimilarity values for the other variables Y_3, Y_4, Y_6, Y_8, Y_{10} that detect the optimal bipartition at the first stage can be obtained as follows:

$$\begin{aligned} Y_3 &: \quad D_{GD}(y'_{53}, y'_{43}) = 0.181, \\ Y_4 &: \quad D_{GD}(y'_{34}, y'_{54}) = 0.260, \\ Y_6 &: \quad D_{GD}(y'_{66}, y'_{56}) = 0.662, \\ Y_8 &: \quad D_{GD}(y'_{48}, y'_{58}) = 0.200, \\ Y_{10} &: \quad D_{GD}(y'_{6,10}, y'_{5,10}) = 0.715. \end{aligned}$$

Since the dissimilarity based on the Y_1 variable has the largest value of 1.223, the binary question for the first stage is based on the variable Y_1, and from Equation (3.15), the cut point can be obtained using the mean for the variable Y_1 of the union of two transformed objects $\mathbf{y}'_6$ and $\mathbf{y}'_5$. That is, the cut point is $M^*_{(6\cup5)1} = 2596.83$. Thus, the binary question for the first stage is 'Is $Y_1 \leq 2596.83$?'. As shown in Figure 5.1(a), if the answer of this question is 'Yes', then the observation goes into cluster $C_1 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$, but if 'No', then it goes into cluster $C_2 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$.

From the result of the first stage, we start the second stage with two clusters $C_1 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ and $C_2 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\})$. In this stage, either of these two clusters is bipartitioned. Similarly to the first stage, we first sort the objects for each cluster by mean values. Using the mean values for the Y_j variable in Table B.3, we can sort the objects in each cluster for the Y_j variable. For example, for the Y_2 variable, the sorted result is $C_1 = \{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}$ and $C_2 = \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1\}$. Similarly, we can sort the objects for the other variables. In this stage, we have to examine five $(= 2 + 3)$ possible biparitions for each variable. Thus, there are a total of 50 possible bipartitions at this stage.

The within-cluster variance values for two clusters $C_1 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$ and $C_2 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ can be found in Table B.4 in columns 3 and 4, respectively. For example, in Table B.4, the third row of the clustering result for variable Y_1 shows the optimal bipartition, to be $\big(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\}\big)$. The $I(C_1^1)$ and $I(C_1^2)$ values (viz., 1.91 and 2.85, respectively) corresponding to the optimal bipartition are the within-cluster variances for the two clusters $C_1 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$ and $C_2 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ at the second stage. That is, at the second stage, $I(C_1) = 1.91$ and $I(C_2) = 2.85$. Using these values, we can calculate the decrement values for each possible bipartition. The results are shown in Table B.5 of Appendix B.4.

From Table B.5, we see that the decrement value, $I(C_u) - I(C_u^1) - I(C_u^2)$, has the largest value of 1.73 when the cluster $C_2 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ is split into $(\{\mathbf{y}_5\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$. That is, the optimal partition at the second stage is $(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$. Since, as shown in Table B.5, all variables except for variable Y_5 (i.e., 9 variables) detect the same optimal bipartition for the second stage, there can be nine possible binary questions at this stage. Thus, to obtain a unique binary question, we examine the extended Gowda-Diday measure values for each variable that detect the optimal bipartition. That is, the variable with the largest dissimilarity value becomes the unique binary question for this stage. For example, for variable Y_3, the optimal bipartition is $(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5\})$ in ascending order by mean values for Y_3. In this case, the two objects on the boundary of the optimal bipartition are $\mathbf{y}_7$ and $\mathbf{y}_5$. That is, the cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ is split into two clusters in the boundary between two objects $\mathbf{y}_7$ and $\mathbf{y}_5$. In this case, the dissimilarity value for variable Y_3 between two transformed histogram-valued objects $\mathbf{y}'_7$ and $\mathbf{y}'_5$, $D_{GD}(y'_{73}, y'_{53})$, should be examined. Thus, we have to examine the dissimilarity values for nine variables except for Y_5 between two objects on the boundary, using Equation (3.21) as follows:

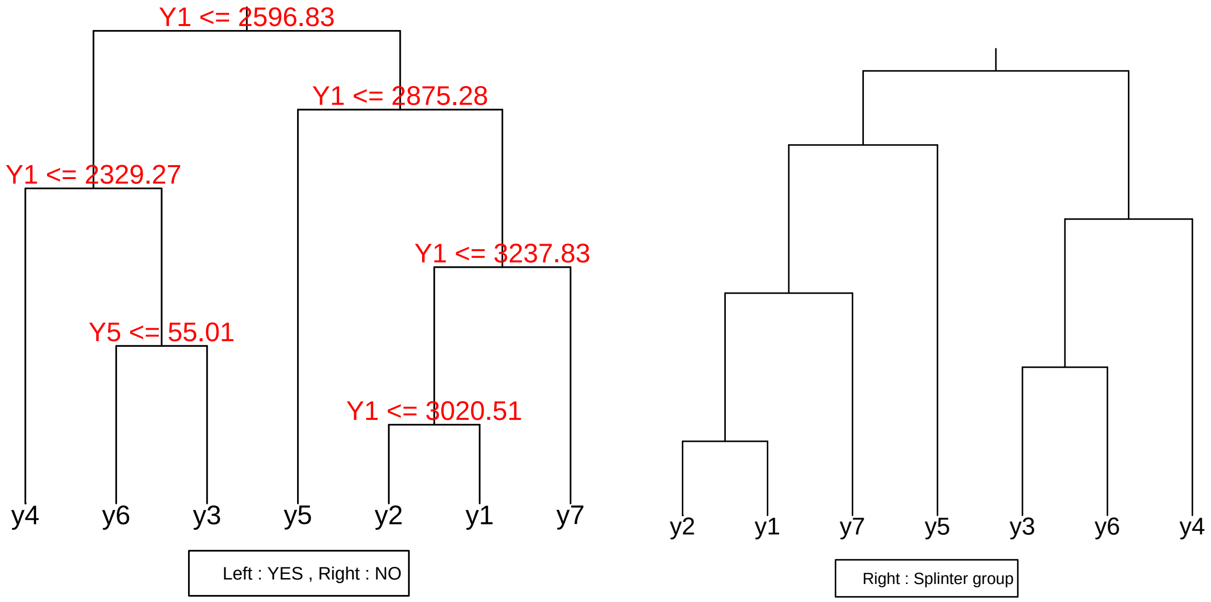
$$\begin{aligned}
Y_1 &: \quad D_{GD}(y'_{51}, y'_{21}) = 0.773, \\
Y_2 &: \quad D_{GD}(y'_{52}, y'_{22}) = 0.225, \\
Y_3 &: \quad D_{GD}(y'_{73}, y'_{53}) = 0.224, \\
Y_4 &: \quad D_{GD}(y'_{54}, y'_{14}) = 0.255, \\
Y_6 &: \quad D_{GD}(y'_{56}, y'_{26}) = 0.701, \\
Y_7 &: \quad D_{GD}(y'_{77}, y'_{57}) = 0.136, \\
Y_8 &: \quad D_{GD}(y'_{58}, y'_{78}) = 0.284, \\
Y_9 &: \quad D_{GD}(y'_{59}, y'_{79}) = 0.320, \\
Y_{10} &: \quad D_{GD}(y'_{5,10}, y'_{1,10}) = 0.355.
\end{aligned}$$

Since the dissimilarity for the variable Y_1 has the largest value of 0.773, the cut point for the second binary question should be the mean value of the union for the variable Y_1 between the two transformed histogram-valued objects $\mathbf{y}'_5$ and $\mathbf{y}'_2$. Thus, from Equation (3.15), the cut point is $M^*_{(5\cup 2)1} = 2875.28$, and the second binary question becomes 'Is $Y_1 \leq 2875.28$?'. As shown in Figure 5.1(a), if the answer for this binary question is 'Yes', the object goes into cluster $\{\mathbf{y}_5\}$. In contrast, if 'No', it goes into cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$.

At the third stage, we start with three clusters ($C_1 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$, $C_2 = \{\mathbf{y}_5\}$, $C_3 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$). One of these three clusters would be bipartitioned, but we do not have to examine cluster C_2 because this cluster has only one object. Thus, in this stage, there are four ($= 2 + 0 + 2$) possible bipartitions for each variable. Similarly to the first and second

Table 5.4: Clustering results for four dissimilarity measures.

Partition	Extended Gowda-Diday (Monothetic and Polythetic)
P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$
P_3	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\})$
P_4	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\})$
P_5	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_7\})$
P_6	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_6\},\ \{\mathbf{y}_7\})$
Partition	Normalized city block (Monothetic and Polythetic)
P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
P_3	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\})$
P_4	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\})$
P_5	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_7\})$
P_6	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_6\},\ \{\mathbf{y}_7\})$
Partition	Normalized Euclidean (Monothetic and Polythetic)
P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$
P_3	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\})$
P_4	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\})$
P_5	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_7\})$
P_6	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_6\},\ \{\mathbf{y}_7\})$
Partition	Normalized CDF (Monothetic and Polythetic)
P_2	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\})$
P_3	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4, \mathbf{y}_5\})$
P_4	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\})$
P_5	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_7\})$
P_6	$(\{\mathbf{y}_1, \mathbf{y}_2\},\ \{\mathbf{y}_3\},\ \{\mathbf{y}_4\},\ \{\mathbf{y}_5\},\ \{\mathbf{y}_6\},\ \{\mathbf{y}_7\})$



(a) Monothetic algorithm

(b) Polythetic algorithm

Figure 5.1: The clustering result for the extended Gowda-Diday dissimilarity measure.

stage, the optimal bipartition and binary question for the third stage can be calculated; and all completed clustering outcomes shown in Table 5.4 and Figure 5.1(a) also can be obtained. Similarly to the case of the extended Gowda-Diday dissimilarity measure of Equation (5.1), clustering results for the normalized city block distance, the normalized Euclidean distance, and the normalized CDF dissimilarity measure of Equation (5.2), (5.3), and (5.4), respectively, can be obtained and are shown in Table 5.4.

Polythetic Method

Let us now apply the polythetic method introduced in Section 3.4.2 to the forest cover type dataset. We start the first stage with $P_1 = (C_1 = \{\mathbf{y}_1, \dots, \mathbf{y}_7\})$. First, we calculate the average weighted dissimilarity, $\bar{D}_u(\mathbf{y}_i)$, for each object using the extended Gowda-Diday dissimilarity matrix of Equation (5.1). For example, suppose that the weight $w_i = 1/n$ for all objects. Then, the average weighted dissimilarity for the object $\mathbf{y}_1$, $\bar{D}_1(\mathbf{y}_1)$, is, from Equation

(3.42),

$$\begin{aligned}\bar{D}_1(\mathbf{y}_1) &= \frac{1}{n(n_1-1)}\sum_{i=1}^{n_1} D_{GD}(\mathbf{y}'_1,\mathbf{y}'_i)\\ &= \frac{1}{7\times 6}(0.000+1.341+6.454+7.305+\cdots+2.953)\\ &= 0.684,\end{aligned}$$

where the dissimilarities $D_{GD}(\mathbf{y}'_1,\mathbf{y}'_i)$ are given in Equation (5.1), and where n is the number of objects in the partition P_1 and n_1 is the number of objects in the cluster C_1.

Similarly, we can obtain the average weighted dissimilarity values for the other objects, and those values are as follows:

$$\begin{aligned}&\bar{D}_1(\mathbf{y}_2)=0.680,\ \bar{D}_1(\mathbf{y}_3)=0.728,\ \bar{D}_1(\mathbf{y}_4)=0.835,\\ &\bar{D}_1(\mathbf{y}_5)=0.660,\ \bar{D}_1(\mathbf{y}_6)=0.752,\ \bar{D}_1(\mathbf{y}_7)=0.783.\end{aligned}$$

From these values, we can obtain the maximum value of the $\bar{D}_1(\mathbf{y}_i),\ i=1,\cdots,7$. Thus, the object $\mathbf{y}_4$ has the maximum average dissimilarity value, and $MAD_1=\max_i\{\bar{D}_1(\mathbf{y}_i),\ i=1,\ldots,7\}=0.835$ (where MAD_1 is the maximum average dissimilarity value for the cluster $C_1=\{\mathbf{y}_1,\ldots,\mathbf{y}_7\}$, see Equation (3.43)). At the first stage, since there is only one cluster, $MMAD=\max_u\{MAD_u\}=MAD_1$ (where $MMAD$ is the maximum average dissimilarity value for all clusters in the current partition, see Equation (3.44)). Therefore, the object $\mathbf{y}_4$ plays a role as a seed, and it goes into the splinter cluster. That is, the splinter cluster $C_1^2=\{\mathbf{y}_4\}\equiv\{\mathbf{y}^*\}$. Then, $C_1^1=\{\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\mathbf{y}_5,\mathbf{y}_6,\mathbf{y}_7\}\equiv\{\mathbf{y}_{(1)},\mathbf{y}_{(2)},\mathbf{y}_{(3)},\mathbf{y}_{(4)},\mathbf{y}_{(5)},\mathbf{y}_{(6)}\}$, where $\mathbf{y}_1\equiv\mathbf{y}_{(1)}$, $\mathbf{y}_2\equiv\mathbf{y}_{(2)}$, $\mathbf{y}_3\equiv\mathbf{y}_{(3)}$, $\mathbf{y}_5\equiv\mathbf{y}_{(4)}$, $\mathbf{y}_6\equiv\mathbf{y}_{(5)}$, and $\mathbf{y}_7\equiv\mathbf{y}_{(6)}$. Note that, unlike the monothetic method, the polythetic method does not consider the order of objects in each cluster.

To obtain the difference of the sums of the within-cluster variances for each object of Equation (3.45), $H_{(i)}$, the within-cluster variance for $C_1^1=\{\mathbf{y}_1,\mathbf{y}_2,\mathbf{y}_3,\mathbf{y}_5,\mathbf{y}_6,\mathbf{y}_7\}$ and for $C_1^2=\{\mathbf{y}_4\}$ are first calculated. Since C_1^2 has only one object at the current status, the within-cluster variance $I(C_1^2)=0$. The within-cluster variance for the cluster C_1^1 is given as, from Equation (3.35),

$$\begin{aligned} I(C_1^1) &= \frac{1}{7 \times 6}\Big\{1.341^2 + 6.454^2 + 4.300^2 + \cdots + 4.770^2 + 7.288^2\}\Big\} \\ &= 9.44. \end{aligned}$$

Let TC^1 and TC^2 be temporary clusters for the clusters C_1^1 and C_1^2, respectively. These temporary clusters play a role as temporary storage corresponding to the clusters C_1^1 and C_1^2. Now, we set $TC^1 = C_1^1$ and $TC^2 = C_1^2$; then $TC^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}$ and $TC^2 = \{\mathbf{y}_4\}$, and one object of TC^1 moves into the cluster TC^2. For example, if the object $\mathbf{y}_{(1)} \equiv \mathbf{y}_1$ moves into the splinter cluster, the cluster $TC^1 = \{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}$ and $TC^2 = \{\mathbf{y}_1, \mathbf{y}_4\}$. Then, the $H_{(1)}$ value can be calculated as follows. From Equation (3.35),

$$\begin{aligned} I(TC^1) &= \frac{1}{7 \times 5}\Big\{6.201^2 + 3.891^2 + 6.285^2 + 3.593^2 + 4.775^2 \\ &\quad +2.077^2 + 7.042^2 + 5.175^2 + 4.770^2 + 7.288^2\}\Big\} = 8.15, \\ I(TC^2) &= \frac{1}{7 \times 2}7.305^2 = 3.81. \end{aligned}$$

Thus, from Equation (3.45),

$$H_{(1)} = (9.44 + 0.00) - (8.15 + 3.81) = -2.52.$$

Similarly, the $H_{(i)}$ values for the other objects can be obtained, and the results are shown in Table B.6 of Appendix B.4.

From Table B.6, when the object $\mathbf{y}_6$ goes into the splinter cluster TC^2, we have the maximum $H_{(i)}$ value. Let MH be the maximum $H_{(i)}$ value (i.e., $MH = \max_i\{H_{(i)},\ i = 1, \ldots, 6\}$); then $MH = 1.41$. Since the MH value is positive, the object $\mathbf{y}_6$ goes into the cluster C_1^2. Thus, currently $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_4, \mathbf{y}_6\}$. The within-cluster variance values for these clusters C_1^1 and C_1^2 are 6.64 and 1.39, respectively. Again, we set temporary clusters where now $TC^1 = C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}$ and $TC^2 = C_1^2 = \{\mathbf{y}_4, \mathbf{y}_6\}$. And then, one object of $TC^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}$ moves into the splinter cluster $TC^2 = \{\mathbf{y}_4, \mathbf{y}_6\}$, and $H_{(i)}$ values are calculated as shown in Table B.7 in Appendix B.4.

From the results of Table B.7, when the object $\mathbf{y}_3$ goes into the splinter cluster TC^2, we have the maximum $H_{(i)}$ value of 3.27. Since $MH > 0$, the object $\mathbf{y}_3$ moves from the cluster C_1^1 into the cluster C_1^2. Thus, now $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$. The within-cluster variance values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ are 2.85 and 2.91, respectively. Similarly to the previous step, we set $TC^1 = C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $TC^2 = C_1^2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$; and then one object of TC^1 goes into the splinter cluster TC^2. Table B.8 (of Appendix B.4) shows this result.

Since $MH = -0.39 < 0$, there is no object to move from the cluster C_1^1 into the cluster C_1^2. Thus, we finally obtain the optimal bipartition for the first stage. That is, the optimal partition for the first stage is $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$. In order to find the optimal bipartition using the polythetic algorithm at this first stage, we examined 15 possible bipartitions. This number of possible bipartitions is much less than that required for the monothetic algorithm (60 possible bipartitions).

We start the second stage with two clusters, $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$. Similarly to the first stage, we have to calculate the average weighted dissimilarity values. To obtain these values, it is more convenient to use the reduced extended Gowda-Diday dissimilarity matrices. The reduced distance/dissimilarity matrix can be obtained from the original distance/dissimilarity matrix corresponding to $\Omega = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. In this case, the original extended Gowda-Diday dissimilarity matrix is given in Equation (5.1). To obtain the reduced dissimilarity matrix corresponding to the cluster $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$, we extract the 1^{st}, 2^{nd}, 5^{th}, and 7^{th} rows and columns from the original dissimilarity matrix of Equation (5.1). Then, the reduced extended Gowda-Diday dissimilarity matrix for the cluster $C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ is given as

$$D_{GD,C_1} = \begin{array}{c} \\ \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_5 \\ \mathbf{y}_7 \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_5 & \mathbf{y}_7 \end{array} \\ \begin{pmatrix} 0.000 & 1.341 & 4.300 & 2.953 \\ 1.341 & 0.000 & 3.891 & 3.593 \\ 4.300 & 3.891 & 0.000 & 4.770 \\ 2.953 & 3.593 & 4.770 & 0.000 \end{pmatrix} \end{array}.$$

Similarly, the reduced dissimilarity matrix for $C_2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ is

$$D_{GD,C_2} = \begin{array}{c} \\ \mathbf{y}_3 \\ \mathbf{y}_4 \\ \mathbf{y}_6 \end{array} \begin{array}{c} \begin{array}{ccc} \mathbf{y}_3 & \mathbf{y}_4 & \mathbf{y}_6 \end{array} \\ \begin{pmatrix} 0.000 & 4.037 & 2.077 \\ 4.037 & 0.000 & 4.416 \\ 2.077 & 4.416 & 0.000 \end{pmatrix} \end{array}.$$

For example, the average weighted dissimilarity value for the objects $\mathbf{y}_1$ and $\mathbf{y}_3$, respectively, are, from Equation (3.35),

$$\begin{aligned} \bar{D}_1(\mathbf{y}_1) &= \frac{1}{7 \times 3}(0.000 + 1.341 + 4.300 + 2.953) \\ &= 0.409, \\ \bar{D}_2(\mathbf{y}_3) &= \frac{1}{7 \times 2}(0.000 + 4.037 + 2.077) \\ &= 0.437. \end{aligned}$$

The other average dissimilarity values for C_1 are $\bar{D}_1(\mathbf{y}_2) = 0.420$, $\bar{D}_1(\mathbf{y}_5) = 0.617$, $\bar{D}_1(\mathbf{y}_7) = 0.539$, and the other average dissimilarity values for C_2 are $\bar{D}_2(\mathbf{y}_4) = 0.604$, $\bar{D}_2(\mathbf{y}_6) = 0.464$. Thus, the maximum value for C_1, from Equation (3.43) is $MAD_1 = 0.617$, and the maximum value for C_2 is $MAD_2 = 0.604$. Finally, the maximum value of MAD_u, $u = 1, 2$ is $MMAD = 0.617$, and the object corresponding to 0.617 is $\mathbf{y}_5$. Thus, $\mathbf{y}_5$ is regarded as a seed, and C_1 is bipartitioned into two clusters at this stage because $\mathbf{y}_5 \in C_1$. That is, $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\} \equiv \{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \mathbf{y}_{(3)}\}$, and $C_1^2 = \{\mathbf{y}_5\} \equiv \{\mathbf{y}^*\}$. The within-cluster variance for C_1^1 is $I(C_1^1) = 1.12$, and for C_1^2 is $I(C_1^2) = 0$.

Now, we set $TC^1 = C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $TC^2 = C_1^2 = \{\mathbf{y}_5\}$; and then $H_{(i)}$ values for each object of C_1 are calculated as shown in Table B.9 of Appendix B.4.

Since the MH value is -0.59 and this value is negative, there is no object to move from the cluster C_1^1 into the cluster C_1^2 at this stage. Thus, the optimal bipartition for the second stage is $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_5\}$, and $P_3 = (C_1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ C_2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ C_3 = \{\mathbf{y}_5\})$. Similarly to the first and second stage, the clustering outcomes for the other stages as shown in Table 5.4 and Figure 5.1(b) can be obtained.

Also, the clustering results based on the normalized city block distance, the normalized Euclidean distance, and the normalized CDF dissimilarity measure of Equation (5.2), (5.3), and (5.4), respectively, can be obtained by the similar procedure to the extended Gowda-Diday matrix of Equation (5.1), and are shown in Figure 5.2, 5.3, and 5.4, respectively.

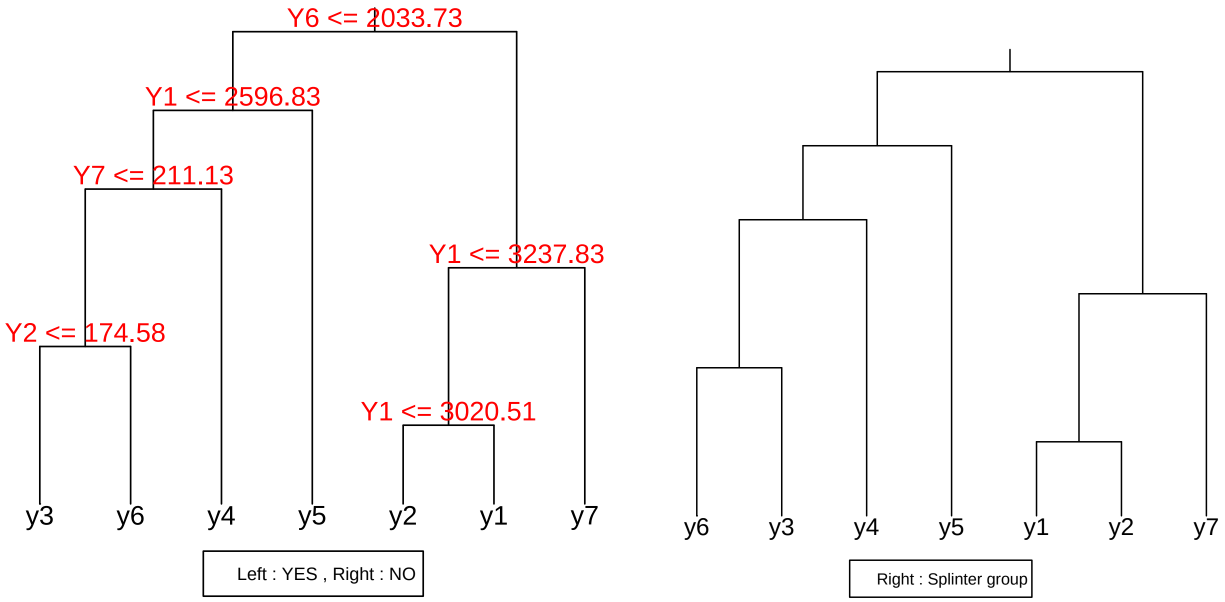
5.3.2 Discussion

Figures 5.1, 5.2, 5.3, 5.4, and Table 5.4 show dendrograms and partitions of the clustering results for each dissimilarity measure and clustering methods. Figure 5.1 gives the dendrograms obtained when the extended Gowda-Diday measure is used; Figure 5.1(a) is obtained when using the monothetic method introduced in Section 3.4.1, and Figure 5.1(b) is given by the polythetic method proposed in Section 3.4.2. Figure 5.2 shows the dendrograms obtained when using the normalized city block distance and both the monothetic and polythetic methods. Figure 5.3 comes from using the normalized Euclidean distance measure, and Figure 5.4 results from using the normalized CDF measure of Equation (3.31). For these dendrograms, the vertical axis represents the r^{th} stage. Also, dendrograms for the monothetic method show the binary questions for each stage. For each binary question, the left side of each node means 'Yes' for binary questions, and the right side represents 'No'. The binary question in the monothetic method identifies the criterion on which bipartitioning is made.

From Figure 5.1(a) obtained by the monothetic algorithm based on the extended Gowda-Diday measure, we see there are six binary questions, with the binary question at the first stage as 'Is $Y_1 \leq 2596.83?'$. This means that bipartitioning is based on variable Y_1, and the cut point for 'Yes' or 'No' is 2596.83. Thus, if the answer is 'Yes', the cover type goes to the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$. Conversely, if 'No', it goes to $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$. That is, if the elevation is equal or less than 2596.83 meters, then the cover type goes to the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$, and if the elevation is larger than 2596.83 meters, then it goes to $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$. For example, suppose that there is a histogram-valued object for a new cover type and the mean for the variable Y_1 of this object, $M_{ij} = M_{i1}$ of Equation (3.13), is 2200 meters. Then, since the mean of this new cover type object is less than the cut point (2596.83), the answer is 'Yes' and this new cover type object is classified into the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$.

At the second stage, the binary question is 'Is $Y_1 \leq 2875.28?'$. Similarly to the first stage, if the answer is 'Yes', then it goes to the cluster $\{\mathbf{y}_5\}$, and if 'No', then it goes to $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$. For example, suppose that we have a histogram-valued object for a new cover type, and the mean for the variable Y_1 of a new histogram-valued object is 2700 meters. Then, the answer for the first binary question, 'Is $Y_1 \leq 2596.83?'$, is 'No', and the answer for the second binary question, 'Is $Y_1 \leq 2875.28?'$, is 'Yes'. Thus, this object is classified into the cluster $\{\mathbf{y}_5\}$. As shown in Figure 5.1(a), all binary questions except for the fifth binary question ('Is $Y_5 \leq 55.01?'$) are related to the variable Y_1 representing the elevation. This means that objects are mainly classified by variable Y_1 in the clustering result obtained using the extended Gowda-Diday measure. That is, the elevation is a important factor to classify cover types when classification is based on the extended Gowda-Diday measure.

As shown in Figure 5.1(b), the polythetic method does not provide binary questions because it uses all p variables to find the optimal bipartition. In the results for the extended Gowda-Diday measure, for all stages, the clustering result of the polythetic method is exactly the same as that of the monothetic method as shown in Figure 5.1(a) and 5.1(b). In Figure 5.1(b), the right side of each node corresponds to the splinter cluster. The polythetic method proposed in Section 3.4.2 starts with finding the object that is the most different from the others within a cluster. That object is called the seed, and the cluster including the seed is called the splinter cluster or group. The polythetic method iteratively compares whether each object is close to a main cluster or a splinter cluster. Thus, from Figure 5.1(b), we know that the splinter cluster is $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ and the object $\mathbf{y}_4$, which is located in the end of the right side in Figure 5.1(b), is the seed at the first stage. At the second stage, the cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ is bipartitioned into $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $\{\mathbf{y}_5\}$. In this case, $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ is the main cluster and $\{\mathbf{y}_5\}$ is the splinter cluster and the seed. In other words, in the cluster $\{\mathbf{y}_1, \dots, \mathbf{y}_7\}$, the object $\mathbf{y}_4$ is the most different from the others, and in the cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$, the object $\mathbf{y}_5$ is the most different in both location and dispersion.



(a) Monothetic algorithm

(b) Polythetic algorithm

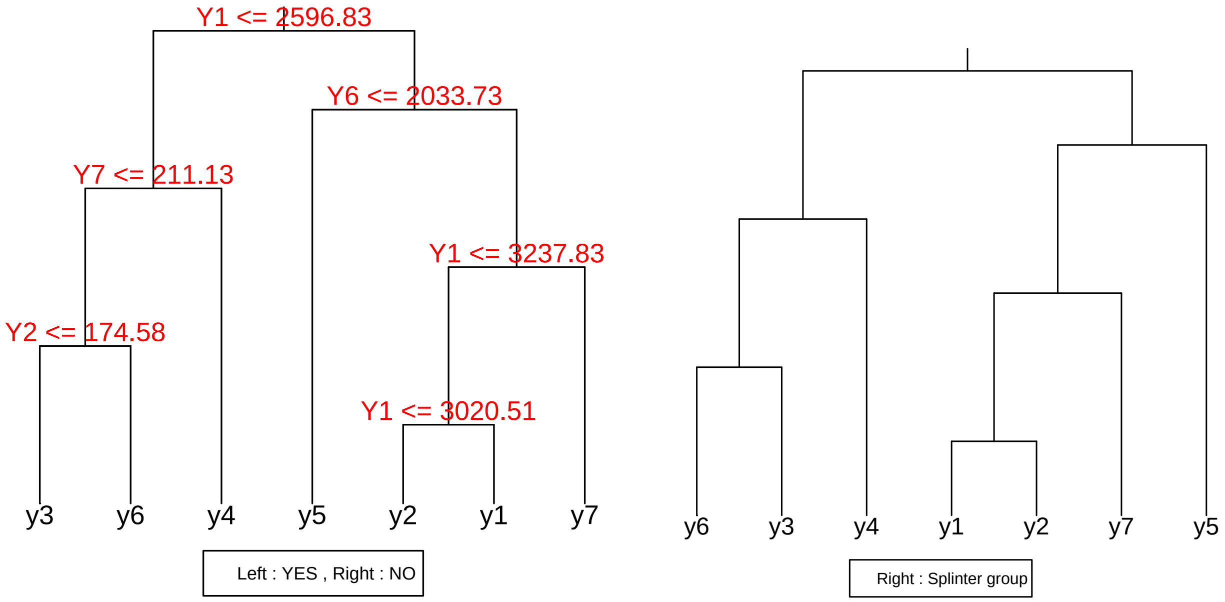
Figure 5.2: The clustering result for the normalized city block distance measure.

Figure 5.2 shows the dendrograms obtained when using the normalized city block distance with $q = 1$ in Equation (3.28). Figure 5.2(a) is obtained using the monothetic clustering method and Figure 5.2(b) uses the polythetic method. The monothetic methods in Figure 5.2(a) provides the binary questions. At the first stage, $P_1 = \{\mathbf{y}_1, \dots, \mathbf{y}_7\}$ is bipartitioned into $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$. This bipartition is a little bit different from the clustering result for the extended Gowda-Diday measure as shown in Figure 5.1(a). The cluster of the object $\mathbf{y}_5$ is changed. However, the third partition of the normalized city block distance is the same as that of the extended Gowda-Diday measure, $P_3 = \big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\}\big)$.

The first binary question is 'Is $Y_6 \leq 2033.73$?'. This means that if the horizontal distance to the nearest roadway is less than 2033.73 meters, then that object belongs to the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$, and otherwise, it goes to $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$. The binary question for the second stage is 'Is $Y_1 \leq 2596.83$?'. The cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$ is bipartitioned into $\big(\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\}\big)$ by the variable Y_1. The six binary questions for this clustering result are related to the variables $Y_1,\ Y_2,\ Y_6$ and Y_7.

From Figure 5.2(b), we know that the splinter cluster for the first stage is $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$. At the second stage, The cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$ is bipartitioned into $\big(\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\}\big)$, and the splinter cluster is $\{\mathbf{y}_5\}$. Similarly to the result for the extended Gowda-Diday measure shown in Figure 5.1, the clustering result for the normalized city block distance by the polythetic algorithm is the same as that by the monothetic algorithm.

Figure 5.3 shows the dendrograms obtained when using the normalized Euclidean distance with $q = 2$ in Equation (3.28). Figure 5.3(a) is obtained using the monothetic clustering algorithm and Figure 5.2(b) comes from the polythetic algorithm. In Figure 5.3(a) by the monothetic method, the first binary question is 'Is $Y_1 \leq 2596.83$?', and $P_1 = \{\mathbf{y}_1, \ldots, \mathbf{y}_7\}$ is bipartitioned into $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ by this binary question. At the second stage, the cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ is bipartitioned into $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_5\}\big)$ by the binary question 'Is $Y_6 \leq 2033.73$?'. For example, suppose that there is a new histogram-valued object, and the mean for Y_1 of this object is 2700 meters and the mean for Y_6 is 2200 meters. Then, this object goes to the cluster $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$. For another example, if the mean for Y_1 of a new object is 2200 meters and the mean for Y_6 is 2000 meters, then this object goes to the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$. That is, this object is not affected by the second binary question 'Is $Y_6 \leq 2033.73$?'. The clustering outcome for the normalized Euclidean distance is the exactly same as that for the extended Gowda-Diday measure, but their binary questions are different. Although monothetic clustering results for each measures can be the same, binary questions for each measure can be different.

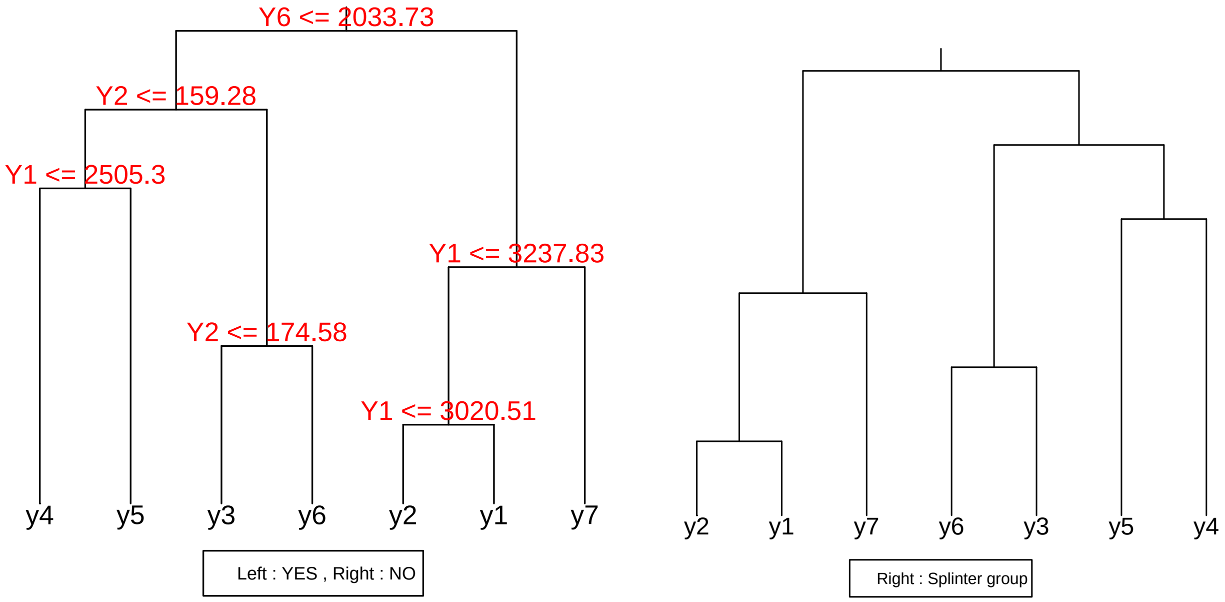


(a) Monothetic algorithm (b) Polythetic algorithm

Figure 5.3: The clustering result for the normalized Euclidean distance measure.

Also, from Figure 5.3, we know that the clustering result for the normalized Euclidean distance by the polythetic algorithm is the same as that by the monothetic algorithm. In Figure 5.3(b), $P_1 = \{\mathbf{y}_1, \ldots, \mathbf{y}_7\}$ is bipartitioned into $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ at the first stage, and the splinter cluster is $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$. This bipartition is the same as that of the extended Gowda-Diday measure shown in Figure 5.1(b). However, the splinter cluster of the extended Gowda-Diday measure is $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ at the first stage, and this splinter cluster is different from that of the normalized Euclidean distance shown in Figure 5.3(b). That is, splinter clusters can be different for different distance/dissimilarity measures.

Figure 5.4 shows the dendrograms obtained when using the normalized CDF dissimilarity measure of Equation (3.31). Figure 5.4(a) is obtained using the monothetic clustering method and Figure 5.4(b) uses the polythetic method. From Figure 5.4(a) by the monothetic method, $P_1 = \{\mathbf{y}_1, \ldots, \mathbf{y}_7\}$ is bipartitioned into $\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$ at the first stage, the binary question for this stage is 'Is $Y_6 \leq 2033.73$?'. This second bipartition P_2 is the same as



(a) Monothetic algorithm

(b) Polythetic algorithm

Figure 5.4: The clustering result for the normalized CDF dissimilarity measure.

that of the normalized city block distance. At the second stage, the cluster $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$ is split into $\{\mathbf{y}_3, \mathbf{y}_6\}$ and $\{\mathbf{y}_4, \mathbf{y}_5\}$ by the binary question 'Is $Y_2 \leq 159.28$?'. That is, if the elevation is less than 2596.83 meters and the aspect is also less than 159.28 azimuth, then the object goes to the cluster $\{\mathbf{y}_4, \mathbf{y}_5\}$. Also, if the elevation is less than 2596.83 meters and the aspect is also larger than 159.28 azimuth, then the object belongs to the cluster $\{\mathbf{y}_3, \mathbf{y}_6\}$. This partition P_3 is different from the third partitions for the other measures. The third partition P_3 for the extended Gowda-Diday, normalized city block and Euclidean measures is $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\}\big)$. However, the partitions $P_4,\ P_5$ and P_6 for all four dissimilarity/distance measures give the same outcomes.

In Figure 5.4(b), the splinter cluster for the first stage is $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}$, and the second stage is $\{\mathbf{y}_4, \mathbf{y}_5\}$. The clustering outcome by the polythetic algorithm is the same as that obtained by the monothetic algorithm.

In summary, Table 5.4 shows partitions of each stage by each distance/dissimilarity measure. For all four distance/dissimilarity measures, the monothetic and polythetic algorithms have the same clustering outcomes at each stage as shown in Table 5.4. The clustering results for the extended Gowda-Diday, normalized city block and Euclidean measures are the same except for the partition P_2 when using the city block distance. The clustering result when using the normalized CDF measure is different from the extended Gowda-Diday and normalized Euclidean measures in P_2 and P_3. The partitions P_2 and P_3 for the extended Gowda-Diday and normalized Euclidean measures are $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}\big)$ and $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\},\ \{\mathbf{y}_5\}\big)$, respectively. However, the partitions P_2 and P_3 are $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_6\}\big)$ and $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_6\},\ \{\mathbf{y}_4, \mathbf{y}_5\}\big)$, respectively. In conclusion, the partitions P_2 and P_3 are different for the four distance/dissimilarity measures, but for all four measures, the partitions P_4, P_5, and P_6 are the same. Thus, all measures give very similar clustering results. Also, we see in Section 5.4 that the optimal number of clusters is $r = 5$.

The cover type dataset includes the cartographic variables such as elevation, aspect, and slope, etc. and all ten variables represent the information for location and circumstance of each cover type. Since the clustering for the cover type dataset is based on these variables, the clustering result is closely related to the areas where the seven cover types are located in. Thus, we verify our clustering result through the the areas where each cover type inhabits.

The cover type data were investigated in four wilderness areas, viz., the Rawah (29,628 hectares), Comanche Peak (27,389 hectares), Neota (3,904 hectares), and Cache la Poudre (3,817 hectares). The proportions and frequencies of each area by types of cover are shown in Table 5.5. In our clustering, we do not consider weights for frequencies of each object. That is, all objects have equal weights. Since all objects have equal weights, proportions of cover types for each area should be considered.

From Table 5.5, Spruce-fir ($\mathbf{y}_1$) and lodgepole pine ($\mathbf{y}_2$) are primarily distributed in the Rawah and Comanche Peak areas. 91.22% of Spruce-fir inhabit the Rawah and Comanche

Table 5.5: Proportion for each cover type and wilderness area.

Symbolic object	Cover Type	Rawah	Neota	Comanche Peak	Cache la Poudre	Total
$\mathbf{y}_1$	Spruce-fir	49.90%	8.78%	41.32%	0.00%	100%
		(105,717)	(18,595)	(87,528)	(0)	(211,840)
$\mathbf{y}_2$	Lodgepole Pine	51.60%	3.17%	44.16%	1.07%	100%
		(146,197)	(8,985)	(125,093)	(3,026)	(283,301)
$\mathbf{y}_3$	Ponderosa Pine	0.00%	0.00%	40.00%	60.00%	100%
		(0)	(0)	(14,300)	(21,454)	(35,754)
$\mathbf{y}_4$	Cottonwood/ Willow	0.00%	0.00%	0.00%	100.00%	100%
		(0)	(0)	(0)	(2,747)	(2,747)
$\mathbf{y}_5$	Aspen	39.83%	0.00%	60.17%	0.00%	100%
		(3,781)	(0)	(5,712)	(0)	(9,493)
$\mathbf{y}_6$	Douglas-fir	0.00%	0.00%	43.91%	56.09%	100%
		(0)	(0)	(7,626)	(9,741)	(17,367)
$\mathbf{y}_7$	Krummholz	24.87%	11.23%	63.90%	0.00%	100%
		(5,101)	(2,304)	(13,105)	(0)	(20,510)

Peak areas, and 95.76% of lodgepole pine are located in these areas. In addition, 49.90% and 41.32% of Spruce-fir are located in the Rawah and Comanche Peak, respectively, and 51.60% and 44.16% of lodgepole pine inhabit the Rawah and Comanche Peak, respectively. Thus, both Spruce-fir and lodgepole pine have similar proportions in the Rawah and Comanche Peak areas. Also, a much lower proportion of these trees is located in the Neota area. Although Krummholz ($\mathbf{y}_7$) is also found in the Rawah, Comanche Peak, and Neota areas, the proportion for the Comanche Peak (63.90%) is much larger than that for the Rawah area (24.87%). Therefore, Spruce-fir and lodgepole pine can be a cluster, and Krummholz can be distinguished from this cluster.

Moreover, ponderosa pine ($\mathbf{y}_3$) and Douglas-fir ($\mathbf{y}_6$) are mainly distributed in the Comanche Peak and Cache la Poudre areas. In addition, 40.00% and 60.00% of ponderosa pine are located in the Comanche Peak and Cache la Poudre, respectively, and 43.91% and 56.09% of Douglas-fir inhabit the Comanche Peak and Cache la Poudre, respectively.

Thus, ponderosa pine and Douglas-fir have a similar distribution across the different areas. Therefore, ponderosa pine and Douglas-fir can be a cluster.

On the contrary, cottonwood/willow ($\mathbf{y}_4$) is only in the Cache la Poudre area, and aspen ($\mathbf{y}_5$) is only found in the Rawah and Comanche Peak. Thus, cottonwood/willow and aspen can be a cluster, respectively. Thus, this result supports the conclusion that there are five clusters in this dataset.

5.4 Diagnostics

For the cover type histogram-valued dataset, processing times for both monothetic and polythetic algorithms (performed using R software, Windows XP with Intel Core2 Duo processor, 3.0GHz, and 2GB RAM) are shown in Table 5.6. These processing times are measured for each distance/dissimilarity measure used, as indicated. As shown in Table 5.6, processing times for each distance/dissimilarity measure are not largely different. The processing time of the monothetic algorithm depends on both the number of objects and the number of variables, and the polythetic algorithm depends on only the number of objects. Thus, the type of distance/dissimilarity measures does not affect processing time for clustering procedure.

Table 5.6: Processing time of clustering algorithms for each distance/dissimilarity measure.

Distance/Dissimilarity	Monothetic	Polythetic
Extended Gowda-Diday	0.0608	0.0084
Normalized city block	0.0607	0.0089
Normalized Euclidean	0.0627	0.0090
Normalized CDF	0.0642	0.0082

In the clustering for the cover type dataset, the processing time of the polythetic algorithm is about seven times faster than that of the monothetic algorithm as shown in Table 5.6. As mentioned in Section 3.4, for the monothetic algorithm, there are $\sum_{u=1}^{r} p(n_u - 1)$ possible bipartitions at the r^{th} stage, where p is the number of variables and n_u is the number of objects in a cluster C_u. On the contrary, the polythetic algorithm has at most $[\{n_r^*(n_r^* - 1)\}/2 - 1]$ possible bipartitions at the r^{th} stage, where n_r^* is the number of objects in the

Table 5.7: Validity index values for each dissimilarity measure and clustering algorithm.

Dissimilarity	Validity	$r = 2$	$r = 3$	$r = 4$	$r = 5$	$r = 6$
Algorithms	Index					
Extended Gowda-Diday	DI_r^s	2.7239	0.9082	1.4355	**3.2053**	2.3972
Mono and Poly	DB_r^s	0.6013	1.4237	0.7699	0.2833	**0.2146**
Normalized city block	DI_r^s	2.7005	1.1711	1.1311	**3.2054**	1.8351
Mono and Poly	DB_r^s	0.4923	1.2427	0.8989	0.3240	**0.2905**
Normalized Euclidean	DI_r^s	3.4310	1.2065	1.0048	**3.5062**	1.3569
Mono and Poly	DB_r^s	0.4363	1.2459	0.9383	**0.3293**	0.3915
Normalized CDF	DI_r^s	**2.8294**	1.5889	1.5583	2.2186	2.1399
Mono and Poly	DB_r^s	0.4336	0.7052	0.6409	0.3769	**0.2684**

splinter cluster. The cover type dataset has ten variables ($p = 10$) and seven objects ($n = 7$). For example, at the first stage, the monothetic method has 60 ($= 10(7 - 1)$) possible bipartitions. In contrast, the polythetic method has at most 20 ($= \{7(7-1)\}/2-1$) possible bipartitions. That is, at the first stage, the number of possible bipartitions of the monothetic method is three times larger than that of the polythetic method. For the clustering result for the extended Gowda-Diday measure, the second partition is $\big(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\},\ \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}\big)$ as shown in Table 5.4. In this case, the monothetic method has 50 possible bipartitions ($= 10(4 - 1) + 10(3 - 1)$). In contrast, since the splinter cluster is $\{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$ as shown in Figure 5.1(b), the polythetic method has at most 5 possible bipartitions ($= \{3(3-1)\}/2-1$). This number is six times smaller than the number of possible biparitions of the monothetic algorithm. Thus, from this dataset, we see that the polythetic method is much faster than the monothetic method when the number of variables is large.

Now, the problem is to find the optimal partition and number of clusters. We investigate these entities using the two cluster validity indexes proposed in Section 4.2. Table 5.7 shows the Dunn and Davis-Bouldin index values for symbolic objects to find the optimal number of clusters. The Dunn and Davis-Bouldin indexes are calculated from Equation (4.8) and

(4.9), respectively. For these data, the monothetic and polythetic algorithms give the same index values for each of the distance/dissimilarity measures and each r value because the clustering outcomes for the monothetic and polythetic methods are exactly the same.

As mentioned in Section 4.2, a larger Dunn index and a smaller Davis-Bouldin index value gives better clustering results. From Table 5.7, the Dunn index, DI_r^s, has the largest value at $r = 5$ for each distance/dissimilarity measure except for the normalized CDF measure, and the Davis-Bouldin index, DB_r^s, has the smallest values at $r = 6$ except for the normalized Euclidean distance. However, the Davis-Bouldin index values for the extended Gowda-Diday, normalized city block and CDF measures have the second smallest value at $r = 5$. Especially, the Davis-Bouldin index values for the extended Gowda-Diday and normalized city block measures at $r = 5$ are similar to the values when $r = 5$ relatively to other values; while both indexes for the normalized Euclidean distance indicate $r = 5$. Thus, we conclude that there are five clusters in this dataset. Since all four distance/dissimilarity measures have the same clustering outcome at $r = 5$, from Table 5.4, the five clusters are {Spruce-fir, Lodgepole pine}, {Ponderosa pine, Douglas-fir}, {Cottonwood/Willow}, {Aspen}, and {Krummholz}.

Although this analysis is just one example, it shows that the distance/dissimilarity measures and clustering algorithms proposed in this study are a viable method for extracting knowledge from large datasets and more important for data that are histogram-valued objects.

Chapter 6

Conclusion and Future Work

Histogram-valued data analytic methodology is one of the methods that deal with numeric variables in huge datasets when we are interested in classes or groups of individuals, rather than individual observations. Also, histogram-valued data are more informative than interval-valued data because they include estimates of the shape and location of the distribution for each class or group, while interval-valued data give only lower and upper limits. Therefore, clustering methodology for histogram-valued data would generally be more precise than that for interval-valued data. However, histogram-valued data are difficult to handle computationally because observations typically have a different number and length of subintervals.

In this study, we propose a transformation for histogram data, as a technique for handling them more easily computationally. From this technique, we developed new distance/dissimilarity measures for histogram-valued data. Since some of the new distance/dissimilarity measures are based on the existing measures for interval-valued data such as the Gowda-Diday and Ichino-Yaguchi measures, distance/dissimilarity values for mixed datasets including multi-valued, interval-valued, and histogram-valued data can now be obtained.

The monothetic algorithm based on a single variable at a time performs poorly in a structure that depends on combinations of variables. For this structure, the polythetic algorithm, introduced in this study, as a new divisive clustering algorithm for symbolic objects is a better approach because it is based on all p variables. Moreover, the polythetic algorithm is much faster than is the monothetic algorithm when the number of variables are large.

One of the important issues in clustering is how to find the optimal number of clusters. For this issue, we propose two validity indexes for symbolic data based on Dunn's index and Davis-Bouldin's index developed for classical data. Also, we showed that these indexes detect well the number of clusters on the Fisher's iris dataset and simulated datasets. We believe that the methodology proposed in this paper can be a useful method for discovering knowledge in large datasets.

Some future work may focus on dissimilarity measures for modal multi-valued data. If dissimilarity measures for modal multi-valued data based on existing measures such as the Gowda-Diday and Ichino-Yaguchi measures can be obtained, in the symbolic context, we might extract more informative knowledge from large datasets. Another future work may be related to supervised learning methods for histogram data such as the $k-$means clustering method. The hierarchical clustering methods including the agglomerative and divisive clustering methods are unsupervised learning methodologies because these methods do not make use of class information about each object during clustering. However, when class information in datasets such as the number of clusters is known, we need clustering methods for symbolic objects that can use this information. Thus, this can be one of the issues for clustering of histogram-valued objects.

Pyramid clustering for histogram-valued data also can be a future work. Pyramid clustering is a agglomerative method, and the difference from usual hierarchical clustering methods is that clusters of a pyramid can overlap. To date, the literature has pyramid methods for classical data (e.g., Bertrand, 1992; and Bertrand and Diday, 1990) and for intervals (Brito, 1995) only, and so far does not provide the pyramid clustering method for histogram-valued data. Thus, this can be one of issues for clustering of symbolic data.

Bibliography

[1] Arroyo, J. and Maté, C. (2009),"Forecasting histogram time series with k-nearest neighbours methods," *International Journal of Forecasting*, 25, 192–207.

[2] Asuncion, A. and Newman, D.J. (2007), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science. Web: http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[3] Berry, M. J. A. and Linoff, G. (1997), *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley and Sons, Berlin.

[4] Bertrand, P. (1992),"Propriétés caractérisations topologiques d'une représentation pyramidale," *Mathématiques et Sciences Humaines*, 117, 5–28.

[5] Bertrand, P. and Diday, E. (1990),"Une généralisation des arbres hiérarchiques: les représentations pyramidales," *Revue de Statistique Appliquée*, 38, 53–78.

[6] Billard, L. and Diday, E. (2003),"From the statistics of data to the statistics of knowledge: Symbolic data analysis," *Journal of the American Statistical Association*, 98, 470–487.

[7] Billard, L. and Diday, E. (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley and Sons, England.

[8] Bock, H.-H. and Diday, E. (Eds.) (2000), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin.

[9] Brito, P. (1995), “Symbolic objects : Order structure and pyramidal clustering”, *Annals of Operations Research*, 55, 277–297.

[10] Cha, S. H. and Srihari, S. N. (2002), “On measuring the distance between histograms,” *Pattern Recognition*, 35, 1355–1370.

[11] Chavent, M. (1998), “A monothetic clustering method,” *Pattern Recognition Letters*, 19, 989–996.

[12] Chavent, M. (2000), “Criterion-based divisive clustering for symbolic data,” In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, (Eds. Bock, H.-H. and Diday, E.), Springer-Verlag, Berlin, 299–311.

[13] Davis, D. L. and Bouldin, D. W. (1979),“A cluster separation measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 224–227.

[14] De Carvalho, F. A. T. (1994), “Proximity coefficients between Boolean symbolic objects,” In: *New Approaches in Classification and Data Analysis, Series: Studies in Classification, Data Analysis, and Knowledge Organisation*, (Eds. Diday, E., Lechevallier, Y., Schader, M., and Bertrand, P.), Springer-Verlag, Berlin, 387–394.

[15] De Carvalho, F. A. T. (1998), “Extension based proximity coefficients between constrained Boolean symbolic objects,” In: *Proceedings of the 5th Conference of the International Federation of Classification Societies (IFCS’96)*, (Eds. Hayashi, C., Ohsumi, N., Yajima, K., Tanaka, Y., Bock, H.-H., and Baba, Y.), Springer-Verlag, Berlin, 370–378.

[16] Diday, E. (1987),“Introduction à l’approche symbolique en analyse des données,” *Première Journées Symbolique-Numérique*, CEREMADE, Université Paris IX, 21–56.

[17] Diday, E. (1992),“From data to knowledge: probabilist objects for a symbolic data analysis,” In: *Proceedings of the 10th Symposium on Computational Statistics*, (Eds. Dodge, Y. and Whittaker, J.), Springer-Verlag, 1, 193–213.

[18] Diday, E. (1995),"Probabilist, possibilist and belief objects for knowledge analysis," *Annals of Operations Research*, 55, 227–276.

[19] DuMouchel, W., Volinsky, C., Johnson, T., Cortes, C. and Pregibon, D. (1999),"Squashing flat files flatter," In: *Proceedings of the 5th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (Eds. Chaudhuri, S., Madigan, D., and Fayyad, U.), ACM Press, 6–15.

[20] Dunn, J. C. (1974),"Well separated clusters and optimal fuzzy partitions," *Journal of Cybernetica*, 4, 95–104.

[21] Edwards, A. W. F. and Cavalli-Sforza, E. L. (1965), "A method for cluster analysis," *Biometrics*, 21, 362–75.

[22] Esposito, F., Malerba, D., Tamma, V. (2000), "Dissimilarity measures for symbolic objects," In: *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data, Series: Studies in Classification, Data Analysis, and Knowledge Organisation*, (Eds. Bock, H.H. and Diday, E.), 15, Springer-Verlag, Berlin, 165-185

[23] Everitt, B. S., Landau, S., and Leese, M. (2001), *Cluster Analysis*, Arnold, London.

[24] Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, Part II, 179–188.

[25] Friedman, H. P. and Rubin, J. (1967), "On some invariant criteria for grouping data," *Journal of the American Statistical Association*, 62, 1159–1178.

[26] Gordon, A. D. (1999), *Classification*, Chapman and Hall, London.

[27] Gowda, K. C. and Diday, E. (1991a), "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, 24, 567–578.

[28] Gowda, K. C. and Diday, E. (1991b), "Unsupervised learning through symbolic clustering," *Pattern Recognition Letters*, 12, 259–264.

[29] Gowda, K.C. and Diday, E. (1992), "Symbolic clustering using a new similarity measure," *IEEE Transactions on Systems, Man, and Cybernetics*, 22, 368–378.

[30] Gowda, K. C. and Ravi, T. V. (1995), "Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity," *Pattern Recognition*, 28, 1277–1282.

[31] Har-even, M. and Brailovsky, V. L. (1995), "Probabilistic validation approach for clustering," *Pattern Recognition*, 16, 1189–1196.

[32] Hartigan, J. A. and Wong, M. A. (1979), "Algorithm AS 136. A k-means clustering algorithm," *Applied Statistics*, 28, 100–108.

[33] Hausdorff, F. (1937), *Set Theory* (translated into English by Aumann, J. R. 1957), Chelsey, New York.

[34] Ichino, M. (1988), "General metrics for mixed features - The Cartesian space theory for pattern recognition," In: *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*, (Ed. Jiang, X.), Pergamon, Oxford, 494–497.

[35] Ichino, M. and Yaguchi, H. (1994), "Generalized Minkowski metrics for mixed feature type data analysis," *IEEE Transactions on Systems, Man, and Cybernetics*, 24, 698–708.

[36] Inatani, S. and Suzuki, E. (2002), "Data squashing for speeding up boosting-based outlier detection," In: *Foundations of Intelligent Systems: Proceedings of the 13th International Symposium, ISMIS 2002*, (Eds. Hacid, M.-S., Ras, Z. W., Zighed, D. A., and Kodratoff, Y.), Springer, Berlin, 601–611.

[37] Irpino, A. and Verde, R. (2006), "A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data," In: *Data Science and Classification, Proceedings of the 10th Conference of the International Federation of Classification Societies (IFCS*

2006), (Eds. Batagelj, V., Bock, H.-H., Ferligoj, A., and Ziberna, A.), Springer, Berlin, 185–192.

[38] Ismail, M. A. and Kamel, M. S. (1989), "Multidimensional data clustering utilizing hybrid search strategies," *Pattern Recognition*, 22, 75–89.

[39] Kaufman, L. and Rousseeuw, P. J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

[40] Kim, M. and Ramakrishna, R. S. (2005), "New indices for cluster validity assessment," *Pattern Recognition Letters*, 26, 2353–2363.

[41] Lance, G. N. and Williams, W. T. (1966), "Computer programs for hierarchical polythetic classification (similarity analyses)," *The Computer Journal*, 9, 60–64.

[42] Lance, G. N. and Williams, W. T. (1968), "Note on a new information statistic classification program," *The Computer Journal*, 11, 195–197.

[43] MacNaughton-Smith, P., Williams, W. T., Dale, M. B., and Mockett, L. G. (1964), "Dissimilarity analysis: A new technique of hierarchical subdivision," *Nature*, 202, 1034–1035.

[44] MacQueen, J. B. (1967), "Some methods for classification and analysis of multivariate observations," In: *Proceedings of the 5th Berkely Symposium on Mathematical Statistics and Probability*, (Eds. Le Cam, L. M. and Neyman, J.), University of California Press, Berkely and Los Angeles, 1, 281–297.

[45] Madigan, D., Raghavan, N., DuMouchel, W., Nason, M., Posse, C. and Ridgeway, G. (2002), "Likelihood-based data squashing: a modeling approach to instance construction," *Data Mining and Knowledge Discovery*, 6, 173–190.

[46] Malerba, D., Esposito, F., Gioviale, V. and Tamma, V. (2001), "Comparing dissimilarity measures for symbolic data analysis," In: *Proceedings of the Joint Conferences on New*

Techniques and Technologies for Statistcs and Exchange of Technology and Know-how (ETK-NTTS 2001), (Eds. Nanopoulos, P. and Wilkinson, D.), European Communities, Rome, 473–481.

[47] Milligan, G. W. and Cooper, M.C. (1985), “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, 50, 159–179.

[48] Nieddu, L. and Rizzi, A. (2003), “Metrics in symbolic data analysis,” In: *New Developments in Classification and Data Analysis: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society*, (Eds. Vichi, M., Monari, P., Mignani, S., and Montanari, A.), Springer, Berlin, 71–78.

[49] Owen, A. (2003), “Data squashing by empirical likelihood,” *Data Mining and Knowledge Discovery*, 7, 103–113.

[50] Ruspini, E. H. (1970), “Numerical methods for fuzzy clustering,” *Information Science*, 2, 319–350.

[51] Strelkov, V. V. (2008), “A new similarity measure for histogram comparison and its application in time series analysis,” *Pattern Recognition Letters*, 29, 1768–1774.

[52] Struyf, A., Hubert, M., and Rousseeuw, P. J. (1997), “Clustering in an object-oriented environment,” *Journal of Statistical Software*, 1. Web: http://www.jstatsoft.org/v01/i04/paper.

[53] Sturges, H. A. (1926), “The choice of a class interval,” *Journal of the American Statistical Association*, 21, 65–66.

[54] Ward, J. H. (1963), “Hierarchical grouping to optimize an objective function,” *Journal of the American Statistical Association*, 58, 238–244.

[55] Williams, W. T. and Lambert, J. M. (1959), “Multivariate methods in plant ecology,” *Journal of Ecology*, 47, 83–101.

[56] Xie, X. L. and Beni, G. A. (1991),“A validity measure for fuzzy clustering,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3, 841–846.

Appendix A

Fisher's Iris Data

A.1 Histogram-valued Data

Table A.1: Histogram-valued data for iris data.

$\mathbf{y}$	Y_1 =Sepal Length	Y_2 =Sepal Width
$\mathbf{y}_1$	$\{[4.4,4.6),0.3;\ [4.6,4.8),0.1;\ [4.8,5.0),0.4;\ [5.0,5.2),0.1;\ [5.2,5.4),0.1\}$	$\{[2.8,3.0),0.2;\ [3.0,3.2),0.3;\ [3.2,3.4),0.2;\ [3.4,3.6),0.2;\ [3.6,3.8),0.0;\ [3.0,3.2),0.1\}$
$\mathbf{y}_2$	$\{[4.0,4.5),0.1;\ [4.5,5.0),0.2;\ [5.0,5.5),0.4;\ [5.5,6.0),0.3\}$	$\{[3.0,3.5),0.4;\ [3.5,4.0),0.5;\ [4.0,4.5),0.1\}$
$\mathbf{y}_3$	$\{[4.6,4.8),0.3;\ [4.8,5.0),0.2;\ [5.0,5.2),0.4;\ [5.2,5.4),0.1\}$	$\{[3.0,3.2),0.2;\ [3.2,3.4),0.5;\ [3.4,3.6),0.2;\ [3.6,3.8),0.1\}$
$\mathbf{y}_4$	$\{[4.4,4.6),0.1;\ [4.6,4.8),0.1;\ [4.8,5.0),0.3;\ [5.0,5.2),0.2;\ [5.2,5.4),0.1;\ [5.4,5.6),0.2\}$	$\{[3.0,3.2),0.4;\ [3.2,3.4),0.2;\ [3.4,3.6),0.2;\ [3.6,3.8),0.0;\ [3.8,4.0),0.0;\ [4.0,4.2),0.2\}$
$\mathbf{y}_5$	$\{[4.4,4.6),0.3;\ [4.6,4.8),0.1;\ [4.8,5.0),0.3;\ [5.0,5.2),0.2;\ [5.2,5.4),0.1\}$	$\{[2.0,2.5),0.1;\ [2.5,3.0),0.1;\ [3.0,3.5),0.5;\ [3.5,4.0),0.3\}$
$\mathbf{y}_6$	$\{[4.5,5.0),0.1;\ [5.0,5.5),0.2;\ [5.5,6.0),0.1;\ [6.0,6.5),0.3;\ [6.5,7.0),0.3\}$	$\{[2.2,2.4),0.2;\ [2.4,2.6),0.0;\ [2.6,2.8),0.3;\ [2.8,3.0),0.1;\ [3.0,3.2),0.3;\ [3.2,3.4),0.1\}$
$\mathbf{y}_7$	$\{[5.0,5.5),0.1;\ [5.5,6.0),0.6;\ [6.0,6.5),0.2;\ [6.5,7.0),0.1\}$	$\{[2.0,2.2),0.3;\ [2.2,2.4),0.0;\ [2.4,2.6),0.1;\ [2.6,2.8),0.1;\ [2.8,3.0),0.4;\ [3.0,3.2),0.1\}$
$\mathbf{y}_8$	$\{[5.6,5.8),0.1;\ [5.8,6.0),0.1;\ [6.0,6.2),0.3;\ [6.2,6.4),0.2;\ [6.4,6.6),0.1;\ [6.6,6.8),0.2\}$	$\{[2.4,2.6),0.4;\ [2.6,2.8),0.2;\ [2.8,3.0),0.2;\ [3.0,3.2),0.2\}$
$\mathbf{y}_9$	$\{[5.4,5.6),0.5;\ [5.6,5.8),0.1;\ [5.8,6.0),0.2;\ [6.0,6.2),0.0;\ [6.2,6.4),0.1;\ [6.4,6.6),0.0;\ [6.6,6.8),0.1\}$	$\{[2.2,2.4),0.3;\ [2.4,2.6),0.1;\ [2.6,2.8),0.2;\ [2.8,3.0),0.2;\ [3.0,3.2),0.1;\ [3.2,3.4),0.1\}$
$\mathbf{y}_{10}$	$\{[5.0,5.2),0.2;\ [5.2,5.4),0.0;\ [5.4,5.6),0.2;\ [5.6,5.8),0.4;\ [5.8,6.0),0.0;\ [6.0,6.2),0.2\}$	$\{[2.2,2.4),0.1;\ [2.4,2.6),0.3;\ [2.6,2.8),0.2;\ [2.8,3.0),0.4\}$
$\mathbf{y}_{11}$	$\{[4.5,5.0),0.1;\ [5.0,5.5),0.0;\ [5.5,6.0),0.1;\ [6.0,6.5),0.3;\ [6.5,7.0),0.1;\ [7.0,7.5),0.3;\ [7.5,8.0),0.1\}$	$\{[2.4,2.6),0.2;\ [2.6,2.8),0.1;\ [2.8,3.0),0.5;\ [3.0,3.2),0.0;\ [3.2,3.4),0.1;\ [3.4,3.6),0.1\}$
$\mathbf{y}_{12}$	$\{[5.5,6.0),0.3;\ [6.0,6.5),0.4;\ [6.5,7.0),0.1;\ [7.0,7.5),0.0;\ [7.5,8.0),0.2\}$	$\{[2.0,2.5),0.2;\ [2.5,3.0),0.5;\ [3.0,3.5),0.2;\ [3.5,4.0),0.1\}$
$\mathbf{y}_{13}$	$\{[5.5,6.0),0.1;\ [6.0,6.5),0.4;\ [6.5,7.0),0.2;\ [7.0,7.5),0.2;\ [7.5,8.0),0.1\}$	$\{[2.7,2.8),0.5;\ [2.8,2.9),0.0;\ [2.9,3.0),0.2;\ [3.0,3.1),0.0;\ [3.1,3.2),0.2;\ [3.2,3.3),0.1\}$
$\mathbf{y}_{14}$	$\{[6.0,6.5),0.6;\ [6.5,7.0),0.1;\ [7.0,7.5),0.1;\ [7.5,8.0),0.2\}$	$\{[2.6,2.8),0.4;\ [2.8,3.0),0.2;\ [3.0,3.2),0.2;\ [3.2,3.4),0.1;\ [3.4,3.6),0.0;\ [3.6,3.8),0.1\}$
$\mathbf{y}_{15}$	$\{[5.8,6.0),0.2;\ [6.0,6.2),0.1;\ [6.2,6.4),0.1;\ [6.4,6.6),0.1;\ [6.6,6.8),0.4;\ [6.8,7.0),0.1\}$	$\{[2.4,2.6),0.1;\ [2.6,2.8),0.1;\ [2.8,3.0),0.3;\ [3.0,3.2),0.3;\ [3.2,3.4),0.2\}$
		(Continued)

$\mathbf{y}$	Y_3 =Petal Length	Y_4 =Petal Width
$\mathbf{y}_1$	$\{[1.3,1.4),0.6;\ [1.4,1.5),0.3;\ [1.5,1.6),0.0;\ [1.6,1.7),0.1\}$	$\{[0.1,0.2),0.8;\ [0.2,0.3),0.1;\ [0.3,0.4),0.1\}$
$\mathbf{y}_2$	$\{[1.1,1.2),0.2;\ [1.2,1.3),0.1;\ [1.3,1.4),0.2;\ [1.4,1.5),0.3;\ [1.5,1.6),0.1;\ [1.6,1.7),0.1\}$	$\{[0.1,0.2),0.5;\ [0.2,0.3),0.3;\ [0.3,0.4),0.2\}$
$\mathbf{y}_3$	$\{[1.0,1.2),0.1;\ [1.2,1.4),0.1;\ [1.4,1.6),0.5;\ [1.6,1.8),0.2;\ [1.8,2.0),0.1\}$	$\{[0.2,0.3),0.7;\ [0.3,0.4),0.2;\ [0.4,0.5),0.1\}$
$\mathbf{y}_4$	$\{[1.2,1.3),0.3;\ [1.3,1.4),0.2;\ [1.4,1.5),0.4;\ [1.5,1.6),0.1\}$	$\{[0.1,0.2),0.9;\ [0.2,0.3),0.0;\ [0.3,0.4),0.1\}$
$\mathbf{y}_5$	$\{[1.3,1.4),0.6;\ [1.4,1.5),0.1;\ [1.5,1.6),0.2;\ [1.6,1.7),0.0;\ [1.7,1.8),0.0;\ [1.8,1.9),0.1\}$	$\{[0.2,0.3),0.8;\ [0.3,0.4),0.1;\ [0.4,0.5),0.0;\ [0.5,0.6),0.1\}$
$\mathbf{y}_6$	$\{[3.0,3.5),0.1;\ [3.5,4.0),0.2;\ [4.0,4.5),0.2;\ [4.5,5.0),0.5\}$	$\{[1.0,1.1),0.1;\ [1.1,1.2),0.0;\ [1.2,1.3),0.3;\ [1.3,1.4),0.2;\ [1.4,1.5),0.3;\ [1.5,1.6),0.1\}$
$\mathbf{y}_7$	$\{[3.4,3.6),0.2;\ [3.6,3.8),0.0;\ [3.8,4.0),0.2;\ [4.0,4.2),0.2;\ [4.2,4.4),0.1;\ [4.4,4.6),0.2;\ [4.6,4.8),0.1\}$	$\{[1.0,1.1),0.4;\ [1.1,1.2),0.0;\ [1.2,1.3),0.1;\ [1.3,1.4),0.2;\ [1.4,1.5),0.3\}$
$\mathbf{y}_8$	$\{[3.5,4.0),0.2;\ [4.0,4.5),0.3;\ [4.5,5.0),0.5\}$	$\{[1.0,1.2),0.2;\ [1.2,1.4),0.4;\ [1.4,1.6),0.2;\ [1.6,1.8),0.2\}$
$\mathbf{y}_9$	$\{[3.6,3.8),0.2;\ [3.8,4.0),0.2;\ [4.0,4.2),0.1;\ [4.2,4.4),0.1;\ [4.4,4.6),0.2;\ [4.6,4.8),0.1;\ [4.8,5.0),0.0;\ [5.0,5.2),0.1\}$	$\{[1.0,1.1),0.2;\ [1.1,1.2),0.1;\ [1.2,1.3),0.3;\ [1.3,1.4),0.0;\ [1.4,1.5),0.2;\ [1.5,1.6),0.2\}$
$\mathbf{y}_{10}$	$\{[3.0,3.5),0.2;\ [3.5,4.0),0.1;\ [4.0,4.5),0.6;\ [4.5,5.0),0.1\}$	$\{[1.0,1.1),0.2;\ [1.1,1.2),0.3;\ [1.2,1.3),0.4;\ [1.3,1.4),0.1\}$
$\mathbf{y}_{11}$	$\{[4.5,5.0),0.1;\ [5.0,5.5),0.1;\ [5.5,6.0),0.5;\ [6.0,6.5),0.2;\ [6.5,7.0),0.1\}$	$\{[1.6,1.8),0.4;\ [1.8,2.0),0.1;\ [2.0,2.2),0.3;\ [2.2,2.4),0.0;\ [2.4,2.6),0.2\}$
$\mathbf{y}_{12}$	$\{[5.0,5.5),0.8;\ [5.5,6.0),0.0;\ [6.0,6.5),0.0;\ [6.5,7.0),0.2\}$	$\{[1.4,1.6),0.1;\ [1.6,1.8),0.1;\ [1.8,2.0),0.3;\ [2.0,2.2),0.2;\ [2.2,2.4),0.3\}$
$\mathbf{y}_{13}$	$\{[4.5,5.0),0.4;\ [5.0,5.5),0.0;\ [5.5,6.0),0.5;\ [6.0,6.5),0.0;\ [6.5,7.0),0.1\}$	$\{[1.6,1.7),0.1;\ [1.7,1.8),0.4;\ [1.8,1.9),0.0;\ [1.9,2.0),0.2;\ [2.0,2.1),0.2;\ [2.1,2.2),0.0;\ [2.2,2.3),0.1\}$
$\mathbf{y}_{14}$	$\{[4.5,5.0),0.1;\ [5.0,5.5),0.3;\ [5.5,6.0),0.3;\ [6.0,6.5),0.3\}$	$\{[1.4,1.6),0.2;\ [1.6,1.8),0.2;\ [1.8,2.0),0.2;\ [2.0,2.2),0.2;\ [2.2,2.4),0.2\}$
$\mathbf{y}_{15}$	$\{[5.0,5.2),0.6;\ [5.2,5.4),0.1;\ [5.4,5.6),0.1;\ [5.6,5.8),0.1;\ [5.8,6.0),0.1\}$	$\{[1.8,1.9),0.3;\ [1.9,2.0),0.1;\ [2.0,2.1),0.0;\ [2.1,2.2),0.0;\ [2.2,2.3),0.4;\ [2.3,2.4),0.1;\ [2.4,2.5),0.1\}$

A.2 Transformed Histogram-valued Data

Table A.2: Transformed histogram-valued data for iris data.

Y_1 Transformed subinterval	Transformed relative frequency														
	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$	$\mathbf{y}'_8$	$\mathbf{y}'_9$	$\mathbf{y}'_{10}$	$\mathbf{y}'_{11}$	$\mathbf{y}'_{12}$	$\mathbf{y}'_{13}$	$\mathbf{y}'_{14}$	$\mathbf{y}'_{15}$
$[4, 4.2)$	0	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0
$[4.2, 4.4)$	0	0.04	0	0	0	0	0	0	0	0	0	0	0	0	0
$[4.4, 4.6)$	0.3	0.06	0	0.1	0.3	0.02	0	0	0	0	0.02	0	0	0	0
$[4.6, 4.8)$	0.1	0.08	0.3	0.1	0.1	0.04	0	0	0	0	0.04	0	0	0	0
$[4.8, 5)$	0.4	0.08	0.2	0.3	0.3	0.04	0	0	0	0	0.04	0	0	0	0
$[5, 5.2)$	0.1	0.16	0.4	0.2	0.2	0.08	0.04	0	0	0.2	0	0	0	0	0
$[5.2, 5.4)$	0.1	0.16	0.1	0.1	0.1	0.08	0.04	0	0	0	0	0	0	0	0
$[5.4, 5.6)$	0	0.14	0	0.2	0	0.06	0.14	0	0.5	0.2	0.02	0.06	0.02	0	0
$[5.6, 5.8)$	0	0.12	0	0	0	0.04	0.24	0.1	0.1	0.4	0.04	0.12	0.04	0	0
$[5.8, 6)$	0	0.12	0	0	0	0.04	0.24	0.2	0.2	0	0.04	0.12	0.04	0	0.2
$[6, 6.2)$	0	0	0	0	0	0.12	0.08	0.2	0	0.2	0.12	0.16	0.16	0.24	0.1
$[6.2, 6.4)$	0	0	0	0	0	0.12	0.08	0.2	0.1	0	0.12	0.16	0.16	0.24	0.1
$[6.4, 6.6)$	0	0	0	0	0	0.12	0.06	0.1	0	0	0.08	0.1	0.12	0.14	0.1
$[6.6, 6.8)$	0	0	0	0	0	0.12	0.04	0.2	0.1	0	0.04	0.04	0.08	0.04	0.4
$[6.8, 7)$	0	0	0	0	0	0.12	0.04	0	0	0	0.04	0.04	0.08	0.04	0.1
$[7, 7.2)$	0	0	0	0	0	0	0	0	0	0	0.12	0	0.08	0.04	0
$[7.2, 7.4)$	0	0	0	0	0	0	0	0	0	0	0.12	0	0.08	0.04	0
$[7.4, 7.6)$	0	0	0	0	0	0	0	0	0	0	0.08	0.04	0.06	0.06	0
$[7.6, 7.8)$	0	0	0	0	0	0	0	0	0	0	0.04	0.08	0.04	0.08	0
$[7.8, 8)$	0	0	0	0	0	0	0	0	0	0	0.04	0.08	0.04	0.08	0

(Continued)

Y_2 Transformed	Transformed relative frequency														
subinterval	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$	$\mathbf{y}'_8$	$\mathbf{y}'_9$	$\mathbf{y}'_{10}$	$\mathbf{y}'_{11}$	$\mathbf{y}'_{12}$	$\mathbf{y}'_{13}$	$\mathbf{y}'_{14}$	$\mathbf{y}'_{15}$
$[2, 2.1)$	0	0	0	0	0.02	0	0.15	0	0	0	0	0.04	0	0	0
$[2.1, 2.2)$	0	0	0	0	0.02	0	0.15	0	0	0	0	0.04	0	0	0
$[2.2, 2.3)$	0	0	0	0	0.02	0.1	0	0	0.15	0.05	0	0.04	0	0	0
$[2.3, 2.4)$	0	0	0	0	0.02	0.1	0	0	0.15	0.05	0	0.04	0	0	0
$[2.4, 2.5)$	0	0	0	0	0.02	0	0.05	0.1	0.05	0.15	0.1	0.04	0	0	0.05
$[2.5, 2.6)$	0	0	0	0	0.02	0	0.05	0.1	0.05	0.15	0.1	0.1	0	0	0.05
$[2.6, 2.7)$	0	0	0	0	0.02	0.15	0.05	0.15	0.1	0.1	0.05	0.1	0	0.2	0.05
$[2.7, 2.8)$	0	0	0	0	0.02	0.15	0.05	0.15	0.1	0.1	0.05	0.1	0.5	0.2	0.05
$[2.8, 2.9)$	0.1	0	0	0	0.02	0.05	0.2	0.2	0.1	0.2	0.25	0.1	0	0.1	0.15
$[2.9, 3)$	0.1	0	0	0	0.02	0.05	0.2	0.2	0.1	0.2	0.25	0.1	0.2	0.1	0.15
$[3, 3.1)$	0.15	0.08	0.1	0.2	0.1	0.15	0.05	0.05	0.05	0	0	0.04	0	0.1	0.15
$[3.1, 3.2)$	0.15	0.08	0.1	0.2	0.1	0.15	0.05	0.05	0.05	0	0	0.04	0.2	0.1	0.15
$[3.2, 3.3)$	0.1	0.08	0.25	0.1	0.1	0.05	0	0	0.05	0	0.05	0.04	0.1	0.05	0.1
$[3.3, 3.4)$	0.1	0.08	0.25	0.1	0.1	0.05	0	0	0.05	0	0.05	0.04	0	0.05	0.1
$[3.4, 3.5)$	0.1	0.08	0.1	0.1	0.1	0	0	0	0	0	0.05	0.04	0	0	0
$[3.5, 3.6)$	0.1	0.1	0.1	0.1	0.06	0	0	0	0	0	0.05	0.02	0	0	0
$[3.6, 3.7)$	0	0.1	0.05	0	0.06	0	0	0	0	0	0	0.02	0	0.05	0
$[3.7, 3.8)$	0	0.1	0.05	0	0.06	0	0	0	0	0	0	0.02	0	0.05	0
$[3.8, 3.9)$	0.05	0.1	0	0	0.06	0	0	0	0	0	0	0.02	0	0	0
$[3.9, 4)$	0.05	0.1	0	0	0.06	0	0	0	0	0	0	0.02	0	0	0
$[4, 4.1)$	0	0.02	0	0.1	0	0	0	0	0	0	0	0	0	0	0
$[4.1, 4.2)$	0	0.02	0	0.1	0	0	0	0	0	0	0	0	0	0	0
$[4.2, 4.3)$	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0
$[4.3, 4.4)$	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0
$[4.4, 4.5)$	0	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0

(Continued)

Y_3 Transformed subinterval	Transformed relative frequency														
	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$	$\mathbf{y}'_8$	$\mathbf{y}'_9$	$\mathbf{y}'_{10}$	$\mathbf{y}'_{11}$	$\mathbf{y}'_{12}$	$\mathbf{y}'_{13}$	$\mathbf{y}'_{14}$	$\mathbf{y}'_{15}$
$[1, 1.1)$	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0
$[1.1, 1.2)$	0	0.2	0.05	0	0	0	0	0	0	0	0	0	0	0	0
$[1.2, 1.3)$	0	0.1	0.05	0.3	0	0	0	0	0	0	0	0	0	0	0
$[1.3, 1.4)$	0.6	0.2	0.05	0.2	0.6	0	0	0	0	0	0	0	0	0	0
$[1.4, 1.5)$	0.3	0.3	0.25	0.4	0.1	0	0	0	0	0	0	0	0	0	0
$[1.5, 1.6)$	0	0.1	0.25	0.1	0.2	0	0	0	0	0	0	0	0	0	0
$[1.6, 1.7)$	0.1	0.1	0.1	0	0	0	0	0	0	0	0	0	0	0	0
$[1.7, 1.8)$	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
$[1.8, 1.9)$	0	0	0.05	0	0.1	0	0	0	0	0	0	0	0	0	0
$[1.9, 2)$	0	0	0.05	0	0	0	0	0	0	0	0	0	0	0	0
$[2, 2.1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.1, 2.2)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.2, 2.3)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.3, 2.4)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.4, 2.5)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.5, 2.6)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.6, 2.7)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.7, 2.8)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.8, 2.9)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[2.9, 3)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[3, 3.1)$	0	0	0	0	0	0.02	0	0	0	0.04	0	0	0	0	0
$[3.1, 3.2)$	0	0	0	0	0	0.02	0	0	0	0.04	0	0	0	0	0
$[3.2, 3.3)$	0	0	0	0	0	0.02	0	0	0	0.04	0	0	0	0	0
$[3.3, 3.4)$	0	0	0	0	0	0.02	0	0	0	0.04	0	0	0	0	0
$[3.4, 3.5)$	0	0	0	0	0	0.02	0.1	0	0	0.04	0	0	0	0	0
$[3.5, 3.6)$	0	0	0	0	0	0.04	0.1	0.04	0	0.02	0	0	0	0	0
$[3.6, 3.7)$	0	0	0	0	0	0.04	0	0.04	0.1	0.02	0	0	0	0	0
$[3.7, 3.8)$	0	0	0	0	0	0.04	0	0.04	0.1	0.02	0	0	0	0	0
$[3.8, 3.9)$	0	0	0	0	0	0.04	0.1	0.04	0.1	0.02	0	0	0	0	0
$[3.9, 4)$	0	0	0	0	0	0.04	0.1	0.04	0.1	0.02	0	0	0	0	0

(Continued)

Y_3 Transformed subinterval	Transformed relative frequency														
	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$	$\mathbf{y}'_8$	$\mathbf{y}'_9$	$\mathbf{y}'_{10}$	$\mathbf{y}'_{11}$	$\mathbf{y}'_{12}$	$\mathbf{y}'_{13}$	$\mathbf{y}'_{14}$	$\mathbf{y}'_{15}$
$[4, 4.1)$	0	0	0	0	0	0.04	0.1	0.06	0.05	0.12	0	0	0	0	0
$[4.1, 4.2)$	0	0	0	0	0	0.04	0.1	0.06	0.05	0.12	0	0	0	0	0
$[4.2, 4.3)$	0	0	0	0	0	0.04	0.05	0.06	0.05	0.12	0	0	0	0	0
$[4.3, 4.4)$	0	0	0	0	0	0.04	0.05	0.06	0.05	0.12	0	0	0	0	0
$[4.4, 4.5)$	0	0	0	0	0	0.04	0.1	0.06	0.1	0.12	0	0	0	0	0
$[4.5, 4.6)$	0	0	0	0	0	0.1	0.1	0.1	0.1	0.02	0.02	0	0.08	0.02	0
$[4.6, 4.7)$	0	0	0	0	0	0.1	0.05	0.1	0.05	0.02	0.02	0	0.08	0.02	0
$[4.7, 4.8)$	0	0	0	0	0	0.1	0.05	0.1	0.05	0.02	0.02	0	0.08	0.02	0
$[4.8, 4.9)$	0	0	0	0	0	0.1	0	0.1	0	0.02	0.02	0	0.08	0.02	0
$[4.9, 5)$	0	0	0	0	0	0.1	0	0.1	0	0.02	0.02	0	0.08	0.02	0
$[5, 5.1)$	0	0	0	0	0	0	0	0	0.05	0	0.02	0.16	0	0.06	0.3
$[5.1, 5.2)$	0	0	0	0	0	0	0	0	0.05	0	0.02	0.16	0	0.06	0.3
$[5.2, 5.3)$	0	0	0	0	0	0	0	0	0	0	0.02	0.16	0	0.06	0.05
$[5.3, 5.4)$	0	0	0	0	0	0	0	0	0	0	0.02	0.16	0	0.06	0.05
$[5.4, 5.5)$	0	0	0	0	0	0	0	0	0	0	0.02	0.16	0	0.06	0.05
$[5.5, 5.6)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.06	0.05
$[5.6, 5.7)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.06	0.05
$[5.7, 5.8)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.06	0.05
$[5.8, 5.9)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.06	0.05
$[5.9, 6)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0.1	0.06	0.05
$[6, 6.1)$	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0.06	0
$[6.1, 6.2)$	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0.06	0
$[6.2, 6.3)$	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0.06	0
$[6.3, 6.4)$	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0.06	0
$[6.4, 6.5)$	0	0	0	0	0	0	0	0	0	0	0.04	0	0	0.06	0
$[6.5, 6.6)$	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0.02	0	0
$[6.6, 6.7)$	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0.02	0	0
$[6.7, 6.8)$	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0.02	0	0
$[6.8, 6.9)$	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0.02	0	0
$[6.9, 7)$	0	0	0	0	0	0	0	0	0	0	0.02	0.04	0.02	0	0
	(Continued)														

Y_4 Transformed	Transformed relative frequency														
subinterval	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$	$\mathbf{y}'_8$	$\mathbf{y}'_9$	$\mathbf{y}'_{10}$	$\mathbf{y}'_{11}$	$\mathbf{y}'_{12}$	$\mathbf{y}'_{13}$	$\mathbf{y}'_{14}$	$\mathbf{y}'_{15}$
$[0.1, 0.2)$	0.8	0.5	0	0.9	0	0	0	0	0	0	0	0	0	0	0
$[0.2, 0.3)$	0.1	0.3	0.7	0	0.8	0	0	0	0	0	0	0	0	0	0
$[0.3, 0.4)$	0.1	0.2	0	0.1	0	0	0	0	0	0	0	0	0	0	0
$[0.4, 0.5)$	0	0	0.1	0	0	0	0	0	0	0	0	0	0	0	0
$[0.5, 0.6)$	0	0	0	0	0.1	0	0	0	0	0	0	0	0	0	0
$[0.6, 0.7)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[0.7, 0.8)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[0.8, 0.9)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[0.9, 1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$[1, 1.1)$	0	0	0	0	0	0.1	0.4	0.1	0.2	0.2	0	0	0	0	0
$[1.1, 1.2)$	0	0	0	0	0	0	0	0.1	0.1	0.3	0	0	0	0	0
$[1.2, 1.3)$	0	0	0	0	0	0.3	0.1	0.2	0.3	0.4	0	0	0	0	0
$[1.3, 1.4)$	0	0	0	0	0	0.2	0.2	0.2	0	0.1	0	0	0	0	0
$[1.4, 1.5)$	0	0	0	0	0	0.3	0.3	0.1	0.2	0	0	0.05	0	0.1	0
$[1.5, 1.6)$	0	0	0	0	0	0.1	0	0.1	0.2	0	0	0.05	0	0.1	0
$[1.6, 1.7)$	0	0	0	0	0	0	0	0.1	0	0	0.2	0.05	0.1	0.1	0
$[1.7, 1.8)$	0	0	0	0	0	0	0	0.1	0	0	0.2	0.05	0.4	0.1	0
$[1.8, 1.9)$	0	0	0	0	0	0	0	0	0	0	0.05	0.15	0	0.1	0.3
$[1.9, 2)$	0	0	0	0	0	0	0	0	0	0	0.05	0.15	0.2	0.1	0.1
$[2, 2.1)$	0	0	0	0	0	0	0	0	0	0	0.15	0.1	0.2	0.1	0
$[2.1, 2.2)$	0	0	0	0	0	0	0	0	0	0	0.15	0.1	0	0.1	0
$[2.2, 2.3)$	0	0	0	0	0	0	0	0	0	0	0	0.15	0.1	0.1	0.4
$[2.3, 2.4)$	0	0	0	0	0	0	0	0	0	0	0	0.15	0	0.1	0.1
$[2.4, 2.5)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0.1
$[2.5, 2.6)$	0	0	0	0	0	0	0	0	0	0	0.1	0	0	0	0

Appendix B

Forestry Cover Type Data

B.1 Histogram-valued Data

Table B.1: Histogram data for forestry cover type data.

$\mathbf{y}$	Y_1=Elevation
$\mathbf{y}_1$	{[2450,2500),0.00004; [2500,2550),0.00070; [2550,2600),0.00169; [2600,2650),0.00285; [2650,2700),0.00441; [2700,2750),0.01005; [2750,2800),0.01385; [2800,2850),0.02272; [2850,2900),0.02643; [2900,2950),0.04330; [2950,3000),0.06893; [3000,3050),0.08873; [3050,3100),0.10630; [3100,3150),0.12242; [3150,3200),0.13729; [3200,3250),0.13883; [3250,3300),0.09751; [3300,3350),0.05452; [3350,3400),0.03426; [3400,3450),0.01391; [3450,3500),0.00518; [3500,3550),0.00335; [3550,3600),0.00145; [3600,3650),0.00095; [3650,3700),0.00034}
$\mathbf{y}_2$	{[2100,2150),0.00001; [2150,2200),0.00015; [2200,2250),0.00025; [2250,2300),0.00049; [2300,2350),0.00067; [2350,2400),0.00083; [2400,2450),0.00180; [2450,2500),0.00522; [2500,2550),0.01775; [2550,2600),0.02688; [2600,2650),0.03552; [2650,2700),0.04405; [2700,2750),0.05751; [2750,2800),0.06760; [2800,2850),0.07280; [2850,2900),0.09077; [2900,2950),0.11580; [2950,3000),0.12553; [3000,3050),0.10131; [3050,3100),0.06574; [3100,3150),0.05454; [3150,3200),0.04528; [3200,3250),0.03369; [3250,3300),0.02581; [3300,3350),0.00765; [3350,3400),0.00210; [3400,3450),0.00026}
$\mathbf{y}_3$	{[1850,1900),0.00064; [1900,1950),0.00579; [1950,2000),0.01454; [2000,2050),0.02937; [2050,2100),0.03890; [2100,2150),0.04548; [2150,2200),0.04964; [2200,2250),0.05415; [2250,2300),0.07020; [2300,2350),0.09384; [2350,2400),0.09034; [2400,2450),0.09526; [2450,2500),0.08393; [2500,2550),0.08010; [2550,2600),0.08757; [2600,2650),0.07227; [2650,2700),0.04262; [2700,2750),0.02327; [2750,2800),0.01390; [2800,2850),0.00646; [2850,2900),0.00171}
$\mathbf{y}_4$	{[1950,2000),0.00364; [2000,2050),0.04951; [2050,2100),0.08227; [2100,2150),0.13324; [2150,2200),0.13069; [2200,2250),0.16199; [2250,2300),0.17692; [2300,2350),0.16163; [2350,2400),0.07863; [2400,2450),0.01311; [2450,2500),0.00692; [2500,2550),0.00146}
$\mathbf{y}_5$	{[2450,2500),0.00748; [2500,2550),0.01243; [2550,2600),0.03697; [2600,2650),0.02950; [2650,2700),0.05667; [2700,2750),0.18466; [2750,2800),0.19404; [2800,2850),0.20931; [2850,2900),0.16022; [2900,2950),0.08712; [2950,3000),0.02096; [3000,3050),0.00063}
	(Continued)

$\mathbf{y}$	Y_1=Elevation
$\mathbf{y}_6$	{[1850,1900),0.00403; [1900,1950),0.01134; [1950,2000),0.01416; [2000,2050),0.01434; [2050,2100),0.02326; [2100,2150),0.03294; [2150,2200),0.02937; [2200,2250),0.03800; [2250,2300),0.05620; [2300,2350),0.08637; [2350,2400),0.11464; [2400,2450),0.13624; [2450,2500),0.11182; [2500,2550),0.08660; [2550,2600),0.07071; [2600,2650),0.06213; [2650,2700),0.04474; [2700,2750),0.03478; [2750,2800),0.01710; [2800,2850),0.00818; [2850,2900),0.00305}
$\mathbf{y}_7$	{[2850,2900),0.00117; [2900,2950),0.00210; [2950,3000),0.00083; [3000,3050),0.00000; [3050,3100),0.00005; [3100,3150),0.00639; [3150,3200),0.02374; [3200,3250),0.08357; [3250,3300),0.13554; [3300,3350),0.18649; [3350,3400),0.26685; [3400,3450),0.17913; [3450,3500),0.05471; [3500,3550),0.01814; [3550,3600),0.01043; [3600,3650),0.00449; [3650,3700),0.00507; [3700,3750),0.00873; [3750,3800),0.00507; [3800,3850),0.00712; [3850,3900),0.00039}
$\mathbf{y}$	Y_2=Aspect
$\mathbf{y}_1$	{[0,20),0.09491; [20,40),0.09745; [40,60),0.09417; [60,80),0.08232; [80,100),0.06844; [100,120),0.05998; [120,140),0.04870; [140,160),0.03875; [160,180),0.03167; [180,200),0.02961; [200,220),0.02648; [220,240),0.02211; [240,260),0.02440; [260,280),0.03581; [280,300),0.04797; [300,320),0.06106; [320,340),0.06893; [340,360),0.06723}
$\mathbf{y}_2$	{[0,20),0.08054; [20,40),0.08259; [40,60),0.08876; [60,80),0.08609; [80,100),0.07682; [100,120),0.06596; [120,140),0.05919; [140,160),0.05030; [160,180),0.04481; [180,200),0.03875; [200,220),0.03840; [220,240),0.03617; [240,260),0.03561; [260,280),0.03350; [280,300),0.03255; [300,320),0.04162; [320,340),0.05220; [340,360),0.05613}
$\mathbf{y}_3$	{[0,20),0.05997; [20,40),0.05599; [40,60),0.05227; [60,80),0.05367; [80,100),0.06704; [100,120),0.07912; [120,140),0.07356; [140,160),0.06041; [160,180),0.05728; [180,200),0.04581; [200,220),0.04181; [220,240),0.03697; [240,260),0.02839; [260,280),0.03119; [280,300),0.04799; [300,320),0.06724; [320,340),0.07557; [340,360),0.06570}
$\mathbf{y}_4$	{[0,20),0.04186; [20,40),0.04951; [40,60),0.06953; [60,80),0.07499; [80,100),0.10193; [100,120),0.17801; [120,140),0.17546; [140,160),0.06225; [160,180),0.02912; [180,200),0.03386; [200,220),0.02039; [220,240),0.01857; [240,260),0.01602; [260,280),0.01383; [280,300),0.01129; [300,320),0.02657; [320,340),0.04077; [340,360),0.03604}
$\mathbf{y}_5$	{[0,20),0.03813; [20,40),0.04814; [40,60),0.09080; [60,80),0.13526; [80,100),0.13599; [100,120),0.09755; [120,140),0.06763; [140,160),0.05783; [160,180),0.05562; [180,200),0.03877; [200,220),0.03234; [220,240),0.02370; [240,260),0.02602; [260,280),0.03582; [280,300),0.02865; [300,320),0.02507; [320,340),0.03150; [340,360),0.03118}
	(Continued)

$\mathbf{y}$	Y_2=Aspect
$\mathbf{y}_6$	{[0,20),0.11948; [20,40),0.11965; [40,60),0.0885; [60,80),0.05614; [80,100),0.04111; [100,120),0.03374; [120,140),0.02263; [140,160),0.01117; [160,180),0.01273; [180,200),0.01140; [200,220),0.01031; [220,240),0.01480; [240,260),0.02240; [260,280),0.03553; [280,300),0.05384; [300,320),0.10745; [320,340),0.14084; [340,360),0.09829}
$\mathbf{y}_7$	{[0,20),0.08947; [20,40),0.07084; [40,60),0.0705; [60,80),0.08279; [80,100),0.10336; [100,120),0.07587; [120,140),0.05661; [140,160),0.05022; [160,180),0.07060; [180,200),0.03784; [200,220),0.02613; [220,240),0.01638; [240,260),0.01253; [260,280),0.01570; [280,300),0.03374; [300,320),0.04924; [320,340),0.05987; [340,360),0.0783}
$\mathbf{y}$	Y_3=Slope
$\mathbf{y}_1$	{[0,5),0.11639; [5,10),0.27713; [10,15),0.28831; [15,20),0.17787; [20,25),0.08703; [25,30),0.03578; [30,35),0.01267; [35,40),0.00340; [40,45),0.00110; [45,50),0.00029; [50,55),0.00003}
$\mathbf{y}_2$	{[0,5),0.11125; [5,10),0.26495; [10,15),0.28129; [15,20),0.18366; [20,25),0.09282; [25,30),0.04280; [30,35),0.01707; [35,40),0.00508; [40,45),0.00072; [45,50),0.00018; [50,55),0.00008; [55,60),0.00006; [60,65),0.00004}
$\mathbf{y}_3$	{[0,5),0.03583; [5,10),0.10718; [10,15),0.16686; [15,20),0.18476; [20,25),0.18504; [25,30),0.16555; [30,35),0.10270; [35,40),0.04274; [40,45),0.00828; [45,50),0.00106}
$\mathbf{y}_4$	{[0,5),0.07463; [5,10),0.17073; [10,15),0.16199; [15,20),0.16600; [20,25),0.17037; [25,30),0.13724; [30,35),0.08664; [35,40),0.02585; [40,45),0.00582; [45,50),0.00073}
$\mathbf{y}_5$	{[0,5),0.07342; [5,10),0.19973; [10,15),0.20573; [15,20),0.20141; [20,25),0.16064; [25,30),0.10060; [30,35),0.05067; [35,40),0.00453; [40,45),0.00105; [45,50),0.00200; [50,55),0.00021}
$\mathbf{y}_6$	{[0,5),0.04008; [5,10),0.11556; [10,15),0.18771; [15,20),0.22520; [20,25),0.21155; [25,30),0.13693; [30,35),0.06783; [35,40),0.01353; [40,45),0.00104; [45,50),0.00029; [50,55),0.00029}
$\mathbf{y}_7$	{[0,5),0.08708; [5,10),0.23974; [10,15),0.31589; [15,20),0.19488; [20,25),0.08406; [25,30),0.03788; [30,35),0.02116; [35,40),0.01365; [40,45),0.00405; [45,50),0.00151; [50,55),0.00010}
	(Continued)

$\mathbf{y}$	Y_4=Horizontal distance to nearest surface water feature
$\mathbf{y}_1$	{[0,50),0.14387; [50,100),0.11947; [100,150),0.11343; [150,200),0.08039; [200,250),0.09279; [250,300),0.08146; [300,350),0.06280; [350,400),0.05730; [400,450),0.05257; [450,500),0.04144; [500,550),0.03380; [550,600),0.03122; [600,650),0.02216; [650,700),0.01867; [700,750),0.01327; [750,800),0.01104; [800,850),0.00854; [850,900),0.00552; [900,950),0.00389; [950,1000),0.00322; [1000,1050),0.00179; [1050,1100),0.00095; [1100,1150),0.00035; [1150,1200),0.00008}
$\mathbf{y}_2$	{[0,50),0.10166; [50,100),0.11049; [100,150),0.11885; [150,200),0.08855; [200,250),0.10710; [250,300),0.09490; [300,350),0.07191; [350,400),0.06516; [400,450),0.05442; [450,500),0.04157; [500,550),0.03371; [550,600),0.02987; [600,650),0.02104; [650,700),0.01645; [700,750),0.01096; [750,800),0.00893; [800,850),0.00691; [850,900),0.00474; [900,950),0.00324; [950,1000),0.00264; [1000,1050),0.00194; [1050,1100),0.00146; [1100,1150),0.00127; [1150,1200),0.00083; [1200,1250),0.00071; [1250,1300),0.00040; [1300,1350),0.00021; [1350,1400),0.00007}
$\mathbf{y}_3$	{[0,50),0.12555; [50,100),0.14762; [100,150),0.14684; [150,200),0.10374; [200,250),0.11943; [250,300),0.10474; [300,350),0.07842; [350,400),0.06108; [400,450),0.04581; [450,500),0.02959; [500,550),0.01770; [550,600),0.01225; [600,650),0.00498; [650,700),0.00210; [700,750),0.00014}
$\mathbf{y}_4$	{[0,50),0.57809; [50,100),0.06807; [100,150),0.06152; [150,200),0.04405; [200,250),0.05752; [250,300),0.05934; [300,350),0.04296; [350,400),0.03859; [400,450),0.02657; [450,500),0.01420; [500,550),0.00837; [550,600),0.00073}
$\mathbf{y}_5$	{[0,100),0.31191; [100,200),0.24471; [200,300),0.20173; [300,400),0.10566; [400,500),0.06320; [500,600),0.03424; [600,700),0.01728; [700,800),0.00874; [800,900),0.00643; [900,1000),0.00358; [1000,1100),0.00253}
$\mathbf{y}_6$	{[0,50),0.22560; [50,100),0.17297; [100,150),0.16088; [150,200),0.10244; [200,250),0.10958; [250,300),0.08862; [300,350),0.05130; [350,400),0.03950; [400,450),0.02447; [450,500),0.01399; [500,550),0.00656; [550,600),0.00294; [600,650),0.00115}
$\mathbf{y}_7$	{[0,100),0.22355; [100,200),0.15978; [200,300),0.14222; [300,400),0.10566; [400,500),0.09615; [500,600),0.08347; [600,700),0.05076; [700,800),0.03603; [800,900),0.02999; [900,1000),0.02579; [1000,1100),0.02589; [1100,1200),0.01599; [1200,1300),0.00458; [1300,1400),0.00015}
	(Continued)

$\mathbf{y}$	Y_5=Vertical distance to nearest surface water feature
$\mathbf{y}_1$	{[-200,-150),0.00001; [-150,-100),0.00094; [-100,-50),0.00573; [-50,0),0.18038; [0,50),0.50471; [50,100),0.18037; [100,150),0.06997; [150,200),0.03492; [200,250),0.01429; [250,300),0.00439; [300,350),0.00257; [350,400),0.00138; [400,450),0.00034}
$\mathbf{y}_2$	{[-200,-150),0.00011; [-150,-100),0.00134; [-100,-50),0.00739; [-50,0),0.13805; [0,50),0.51084; [50,100),0.20823; [100,150),0.07549; [150,200),0.03343; [200,250),0.01469; [250,300),0.00654; [300,350),0.00244; [350,400),0.00071; [400,450),0.00020; [450,500),0.00013; [500,550),0.00019; [550,600),0.00020}
$\mathbf{y}_3$	{[-140,-120),0.00020; [-120,-100),0.00103; [-100,-80),0.00190; [-80,-60),0.00266; [-60,-40),0.00411; [-40,-20),0.01085; [-20,0),0.08077; [0,20),0.17053; [20,40),0.16367; [40,60),0.13179; [60,80),0.11425; [80,100),0.09112; [100,120),0.06696; [120,140),0.04794; [140,160),0.03672; [160,180),0.02646; [180,200),0.01863; [200,220),0.01404; [220,240),0.00856; [240,260),0.00492; [260,280),0.00221; [280,300),0.00059; [300,320),0.00008}
$\mathbf{y}_4$	{[-40,-20),0.00109; [-20,0),0.44230; [0,20),0.15690; [20,40),0.06480; [40,60),0.05461; [60,80),0.05096; [80,100),0.04223; [100,120),0.04077; [120,140),0.04623; [140,160),0.04368; [160,180),0.02621; [180,200),0.01493; [200,220),0.00655; [220,240),0.00400; [240,260),0.00364; [260,280),0.00109}
$\mathbf{y}_5$	{[-140,-120),0.00021; [-120,-100),0.00158; [-100,-80),0.00221; [-80,-60),0.00411; [-60,-40),0.00569; [-40,-20),0.01148; [-20,0),0.15348; [0,20),0.21964; [20,40),0.13305; [40,60),0.11429; [60,80),0.09976; [80,100),0.07806; [100,120),0.05320; [120,140),0.03676; [140,160),0.02697; [160,180),0.01917; [180,200),0.01496; [200,220),0.01243; [220,240),0.00885; [240,260),0.00369; [260,280),0.00042}
$\mathbf{y}_6$	{[-140,-120),0.00017; [-120,-100),0.00035; [-100,-80),0.00063; [-80,-60),0.00161; [-60,-40),0.00225; [-40,-20),0.00726; [-20,0),0.14890; [0,20),0.22595; [20,40),0.16773; [40,60),0.12570; [60,80),0.10641; [80,100),0.08159; [100,120),0.05672; [120,140),0.03104; [140,160),0.01825; [160,180),0.01198; [180,200),0.00662; [200,220),0.00392; [220,240),0.00167; [240,260),0.00069; [260,280),0.00046; [280,300),0.00012}
$\mathbf{y}_7$	{[-100,-50),0.00463; [-50,0),0.15241; [0,50),0.38425; [50,100),0.18893; [100,150),0.11224; [150,200),0.07518; [200,250),0.04247; [250,300),0.02199; [300,350),0.01195; [350,400),0.00551; [400,450),0.00044}
	(Continued)

$\mathbf{y}$	Y_6=Horizontal distance to nearest roadway
$\mathbf{y}_1$	{[0,500),0.04487; [500,1000),0.09847; [1000,1500),0.13361; [1500,2000),0.12812; [2000,2500),0.12070; [2500,3000),0.10438; [3000,3500),0.09860; [3500,4000),0.07088; [4000,4500),0.06256; [4500,5000),0.04972; [5000,5500),0.04307; [5500,6000),0.03309; [6000,6500),0.01163; [6500,7000),0.00032}
$\mathbf{y}_2$	{[0,500),0.06916; [500,1000),0.13858; [1000,1500),0.15308; [1500,2000),0.13071; [2000,2500),0.10419; [2500,3000),0.09182; [3000,3500),0.07236; [3500,4000),0.05029; [4000,4500),0.04656; [4500,5000),0.03702; [5000,5500),0.04006; [5500,6000),0.04452; [6000,6500),0.01731; [6500,7000),0.00415; [7000,7500),0.00019}
$\mathbf{y}_3$	{[0,200),0.07230; [200,400),0.13131; [400,600),0.14650; [600,800),0.11884; [800,1000),0.12345; [1000,1200),0.11395; [1200,1400),0.09571; [1400,1600),0.06567; [1600,1800),0.03591; [1800,2000),0.02472; [2000,2200),0.02268; [2200,2400),0.01863; [2400,2600),0.01390; [2600,2800),0.00825; [2800,3000),0.00369; [3000,3200),0.00299; [3200,3400),0.00140; [3400,3600),0.00008}
$\mathbf{y}_4$	{[0,100),0.00364; [100,200),0.01820; [200,300),0.03531; [300,400),0.05169; [400,500),0.06261; [500,600),0.06443; [600,700),0.06553; [700,800),0.07353; [800,900),0.08373; [900,1000),0.08810; [1000,1100),0.07754; [1100,1200),0.11067; [1200,1300),0.10339; [1300,1400),0.08482; [1400,1500),0.05461; [1500,1600),0.01529; [1600,1700),0.00655; [1700,1800),0.00036}
$\mathbf{y}_5$	{[0,500),0.26261; [500,1000),0.16991; [1000,1500),0.14189; [1500,2000),0.19741; [2000,2500),0.14284; [2500,3000),0.03877; [3000,3500),0.01001; [3500,4000),0.00000; [4000,4500),0.00000; [4500,5000),0.02718; [5000,5500),0.00938}
$\mathbf{y}_6$	{[0,200),0.03576; [200,400),0.09247; [400,600),0.12334; [600,800),0.14585; [800,1000),0.12201; [1000,1200),0.12161; [1200,1400),0.11205; [1400,1600),0.07802; [1600,1800),0.05326; [1800,2000),0.04284; [2000,2200),0.03985; [2200,2400),0.02136; [2400,2600),0.00535; [2600,2800),0.00213; [2800,3000),0.00311; [3000,3200),0.00098}
$\mathbf{y}_7$	{[0,500),0.00073; [500,1000),0.07138; [1000,1500),0.10478; [1500,2000),0.13915; [2000,2500),0.13725; [2500,3000),0.14222; [3000,3500),0.10731; [3500,4000),0.11355; [4000,4500),0.11024; [4500,5000),0.03613; [5000,5500),0.03725}
	(Continued)

y	Y_7=A relative measure of incident sunlight at 09:00A.M. on the summer solstice
$\mathbf{y}_1$	{[0,10),0.00002; [10,20),0.00000; [20,30),0.00000; [30,40),0.00000; [40,50),0.00000; [50,60),0.00005; [60,70),0.00009; [70,80),0.00027; [80,90),0.00043; [90,100),0.00065; [100,110),0.00107; [110,120),0.00222; [120,130),0.00417; [130,140),0.00645; [140,150),0.00895; [150,160),0.01431; [160,170),0.02378; [170,180),0.04274; [180,190),0.06667; [190,200),0.09595; [200,210),0.13396; [210,220),0.17491; [220,230),0.18921; [230,240),0.14581; [240,250),0.07561; [250,260),0.01265}
$\mathbf{y}_2$	{[0,10),0.00002; [10,20),0.00000; [20,30),0.00000; [30,40),0.00000; [40,50),0.00000; [50,60),0.00005; [60,70),0.00006; [70,80),0.00018; [80,90),0.00034; [90,100),0.00073; [100,110),0.00120; [110,120),0.00182; [120,130),0.00323; [130,140),0.00566; [140,150),0.00998; [150,160),0.01522; [160,170),0.02335; [170,180),0.03966; [180,190),0.05873; [190,200),0.08672; [200,210),0.11880; [210,220),0.16033; [220,230),0.19822; [230,240),0.17427; [240,250),0.08956; [250,260),0.01187}
$\mathbf{y}_3$	{[40,50),0.00003; [50,60),0.00011; [60,70),0.00078; [70,80),0.00159; [80,90),0.00372; [90,100),0.00554; [100,110),0.01189; [110,120),0.01673; [120,130),0.02363; [130,140),0.03370; [140,150),0.04128; [150,160),0.04570; [160,170),0.05166; [170,180),0.05253; [180,190),0.05697; [190,200),0.06699; [200,210),0.07129; [210,220),0.08318; [220,230),0.11448; [230,240),0.13232; [240,250),0.13464; [250,260),0.05124}
$\mathbf{y}_4$	{[120,130),0.00036; [130,140),0.00291; [140,150),0.00364; [150,160),0.00837; [160,170),0.01638; [170,180),0.02330; [180,190),0.03058; [190,200),0.05315; [200,210),0.06516; [210,220),0.10885; [220,230),0.13578; [230,240),0.14161; [240,250),0.21478; [250,260),0.19512}
$\mathbf{y}_5$	{[120,130),0.00042; [130,140),0.00200; [140,150),0.00348; [150,160),0.00737; [160,170),0.01464; [170,180),0.02770; [180,190),0.04761; [190,200),0.06447; [200,210),0.07405; [210,220),0.11472; [220,230),0.18951; [230,240),0.19277; [240,250),0.20299; [250,260),0.05825}
$\mathbf{y}_6$	{[0,20),0.00017; [20,40),0.00006; [40,60),0.00012; [60,80),0.00023; [80,100),0.00115; [100,120),0.01457; [120,140),0.05551; [140,160),0.12006; [160,180),0.16704; [180,200),0.18426; [200,220),0.20856; [220,240),0.19180; [240,260),0.05649}
$\mathbf{y}_7$	{[80,90),0.00005; [90,100),0.00010; [100,110),0.00083; [110,120),0.00122; [120,130),0.00141; [130,140),0.00254; [140,150),0.00531; [150,160),0.01009; [160,170),0.01633; [170,180),0.03091; [180,190),0.06724; [190,200),0.10078; [200,210),0.10843; [210,220),0.13993; [220,230),0.19430; [230,240),0.16568; [240,250),0.12750; [250,260),0.02735}
	(Continued)

$\mathbf{y}$	Y_8=A relative measure of incident sunlight at noon on the summer solstice
$\mathbf{y}_1$	{[70,80),0.00001; [80,90),0.00003; [90,100),0.00007; [100,110),0.00011; [110,120),0.00012; [120,130),0.00028; [130,140),0.00058; [140,150),0.00133; [150,160),0.00278; [160,170),0.00576; [170,180),0.01282; [180,190),0.02866; [190,200),0.05254; [200,210),0.10000; [210,220),0.17525; [220,230),0.23468; [230,240),0.21813; [240,250),0.13426; [250,260),0.03260}
$\mathbf{y}_2$	{[0,10),0.00002; [10,20),0.00000; [20,30),0.00000; [30,40),0.00000; [40,50),0.00001; [50,60),0.00001; [60,70),0.00001; [70,80),0.00001; [80,90),0.00002; [90,100),0.00004; [100,110),0.00006; [110,120),0.00012; [120,130),0.00024; [130,140),0.00040; [140,150),0.00096; [150,160),0.00253; [160,170),0.00588; [170,180),0.01258; [180,190),0.02334; [190,200),0.04702; [200,210),0.09411; [210,220),0.16101; [220,230),0.22437; [230,240),0.20600; [240,250),0.16409; [250,260),0.05717}
$\mathbf{y}_3$	{[90,100),0.00011; [100,110),0.00039; [110,120),0.00078; [120,130),0.00336; [130,140),0.00722; [140,150),0.01491; [150,160),0.02179; [160,170),0.03295; [170,180),0.04416; [180,190),0.05474; [190,200),0.07585; [200,210),0.10667; [210,220),0.13260; [220,230),0.14521; [230,240),0.14706; [240,250),0.14605; [250,260),0.06615}
$\mathbf{y}_4$	{[130,140),0.00036; [140,150),0.00218; [150,160),0.00728; [160,170),0.01820; [170,180),0.03203; [180,190),0.06079; [190,200),0.08518; [200,210),0.13287; [210,220),0.17510; [220,230),0.22024; [230,240),0.13760; [240,250),0.09028; [250,260),0.03786}
$\mathbf{y}_5$	{[90,100),0.00084; [100,110),0.00053; [110,120),0.00126; [120,130),0.00053; [130,140),0.00053; [140,150),0.00158; [150,160),0.01338; [160,170),0.03424; [170,180),0.04245; [180,190),0.05594; [190,200),0.06426; [200,210),0.08311; [210,220),0.13547; [220,230),0.18224; [230,240),0.17013; [240,250),0.16402; [250,260),0.04951}
$\mathbf{y}_6$	{[90,100),0.00006; [100,110),0.00012; [110,120),0.00046; [120,130),0.00063; [130,140),0.00432; [140,150),0.01296; [150,160),0.02603; [160,170),0.03501; [170,180),0.04629; [180,190),0.07186; [190,200),0.10998; [200,210),0.14919; [210,220),0.18293; [220,230),0.16065; [230,240),0.10520; [240,250),0.07474; [250,260),0.01958}
$\mathbf{y}_7$	{[90,100),0.00020; [100,110),0.00029; [110,120),0.00088; [120,130),0.00200; [130,140),0.00219; [140,150),0.00293; [150,160),0.00497; [160,170),0.00878; [170,180),0.01531; [180,190),0.02477; [190,200),0.04993; [200,210),0.12335; [210,220),0.19083; [220,230),0.21312; [230,240),0.18679; [240,250),0.15110; [250,260),0.02257}
	(Continued)

$\mathbf{y}$	Y_9=A relative measure of incident sunlight at 03:00P.M. on the summer solstice
$\mathbf{y}_1$	{[0,10),0.00192; [10,20),0.00048; [20,30),0.00108; [30,40),0.00220; [40,50),0.00396; [50,60),0.00718; [60,70),0.01118; [70,80),0.01809; [80,90),0.02600; [90,100),0.03790; [100,110),0.05530; [110,120),0.07487; [120,130),0.09748; [130,140),0.11494; [140,150),0.12035; [150,160),0.10469; [160,170),0.08951; [170,180),0.07779; [180,190),0.06034; [190,200),0.04186; [200,210),0.02637; [210,220),0.01489; [220,230),0.00765; [230,240),0.00321; [240,250),0.00074; [250,260),0.00002}
$\mathbf{y}_2$	{[0,10),0.00166; [10,20),0.00071; [20,30),0.00151; [30,40),0.00259; [40,50),0.00374; [50,60),0.00624; [60,70),0.00987; [70,80),0.01631; [80,90),0.02460; [90,100),0.04025; [100,110),0.06290; [110,120),0.08319; [120,130),0.10640; [130,140),0.11908; [140,150),0.11900; [150,160),0.09753; [160,170),0.08047; [170,180),0.06765; [180,190),0.05760; [190,200),0.04247; [200,210),0.02662; [210,220),0.01491; [220,230),0.00832; [230,240),0.00466; [240,250),0.00159; [250,260),0.00012}
$\mathbf{y}_3$	{[0,20),0.01222; [20,40),0.02036; [40,60),0.04072; [60,80),0.06830; [80,100),0.08942; [100,120),0.11993; [120,140),0.14214; [140,160),0.13716; [160,180),0.11878; [180,200),0.10751; [200,220),0.08528; [220,240),0.05082; [240,260),0.00736}
$\mathbf{y}_4$	{[0,20),0.03240; [20,40),0.04878; [40,60),0.08336; [60,80),0.12887; [80,100),0.13287; [100,120),0.12523; [120,140),0.13688; [140,160),0.13724; [160,180),0.09902; [180,200),0.04878; [200,220),0.01929; [220,240),0.00728}
$\mathbf{y}_5$	{[0,20),0.01949; [20,40),0.04667; [40,60),0.06342; [60,80),0.09818; [80,100),0.10060; [100,120),0.11977; [120,140),0.16876; [140,160),0.15548; [160,180),0.10050; [180,200),0.08501; [200,220),0.03655; [220,240),0.00558}
$\mathbf{y}_6$	{[0,20),0.00029; [20,40),0.00846; [40,60),0.02223; [60,80),0.05038; [80,100),0.08315; [100,120),0.11505; [120,140),0.14833; [140,160),0.14706; [160,180),0.14695; [180,200),0.13906; [200,220),0.10336; [220,240),0.03512; [240,260),0.00058}
$\mathbf{y}_7$	{[0,20),0.01560; [20,40),0.00941; [40,60),0.01760; [60,80),0.04276; [80,100),0.08108; [100,120),0.14603; [120,140),0.18952; [140,160),0.24334; [160,180),0.15553; [180,200),0.07782; [200,220),0.01897; [220,240),0.00234}
	(Continued)

$\mathbf{y}$	Y_{10}=Horizontal distance to nearest historic wildfire ignition point
$\mathbf{y}_1$	{[0,500),0.06696; [500,1000),0.15156; [1000,1500),0.17071; [1500,2000),0.16847; [2000,2500),0.15370; [2500,3000),0.11807; [3000,3500),0.06625; [3500,4000),0.03612; [4000,4500),0.02021; [4500,5000),0.01444; [5000,5500),0.01250; [5500,6000),0.00927; [6000,6500),0.00782; [6500,7000),0.00372; [7000,7500),0.00021}
$\mathbf{y}_2$	{[0,500),0.04930; [500,1000),0.13965; [1000,1500),0.18532; [1500,2000),0.17819; [2000,2500),0.15398; [2500,3000),0.10811; [3000,3500),0.03936; [3500,4000),0.02231; [4000,4500),0.02521; [4500,5000),0.02623; [5000,5500),0.02261; [5500,6000),0.02341; [6000,6500),0.02176; [6500,7000),0.00434; [7000,7500),0.00022}
$\mathbf{y}_3$	{[0,200),0.03658; [200,400),0.12032; [400,600),0.16658; [600,800),0.15939; [800,1000),0.15044; [1000,1200),0.11663; [1200,1400),0.08824; [1400,1600),0.05552; [1600,1800),0.03801; [1800,2000),0.02495; [2000,2200),0.01233; [2200,2400),0.01404; [2400,2600),0.01133; [2600,2800),0.00464; [2800,3000),0.00098}
$\mathbf{y}_4$	{[0,200),0.05934; [200,400),0.14234; [400,600),0.15399; [600,800),0.14379; [800,1000),0.14671; [1000,1200),0.08118; [1200,1400),0.10339; [1400,1600),0.07972; [1600,1800),0.06007; [1800,2000),0.02949}
$\mathbf{y}_5$	{[0,500),0.06352; [500,1000),0.22891; [1000,1500),0.22332; [1500,2000),0.24302; [2000,2500),0.17402; [2500,3000),0.02876; [3000,3500),0.00421; [3500,4000),0.00221; [4000,4500),0.00000; [4500,5000),0.00000; [5000,5500),0.00790; [5500,6000),0.00938; [6000,6500),0.01475}
$\mathbf{y}_6$	{[0,200),0.02148; [200,400),0.07572; [400,600),0.11839; [600,800),0.16946; [800,1000),0.15679; [1000,1200),0.11816; [1200,1400),0.11338; [1400,1600),0.07071; [1600,1800),0.03921; [1800,2000),0.03340; [2000,2200),0.02821; [2200,2400),0.01831; [2400,2600),0.01889; [2600,2800),0.01382; [2800,3000),0.00409}
$\mathbf{y}_7$	{[0,500),0.06080; [500,1000),0.13774; [1000,1500),0.13876; [1500,2000),0.17367; [2000,2500),0.15095; [2500,3000),0.10922; [3000,3500),0.09274; [3500,4000),0.09473; [4000,4500),0.04096; [4500,5000),0.00044}

B.2 Transformed Histogram-valued Data

Table B.2: Transformed histogram-valued data for forestry cover type data.

Y_1	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{1k}, b_{1,k+1})$	p'_{11k}	p'_{21k}	p'_{31k}	p'_{41k}	p'_{51k}	p'_{61k}	p'_{71k}
$[1850, 1900)$	0	0	0.00064	0	0	0.00403	0
$[1900, 1950)$	0	0	0.00579	0	0	0.01134	0
$[1950, 2000)$	0	0	0.01454	0.00364	0	0.01416	0
$[2000, 2050)$	0	0	0.02937	0.04951	0	0.01434	0
$[2050, 2100)$	0	0	0.0389	0.08227	0	0.02326	0
$[2100, 2150)$	0	0	0.04548	0.13324	0	0.03294	0
$[2150, 2200)$	0	0.00015	0.04964	0.13069	0	0.02937	0
$[2200, 2250)$	0	0.00025	0.05415	0.16199	0	0.038	0
$[2250, 2300)$	0	0.00049	0.0702	0.17692	0	0.0562	0
$[2300, 2350)$	0	0.00067	0.09384	0.16163	0	0.08637	0
$[2350, 2400)$	0	0.00083	0.09034	0.07863	0	0.11464	0
$[2400, 2450)$	0	0.0018	0.09526	0.01311	0	0.13624	0
$[2450, 2500)$	0	0.00522	0.08393	0.00692	0.00748	0.11182	0
$[2500, 2550)$	0.0007	0.01775	0.0801	0.00146	0.01243	0.0866	0
$[2550, 2600)$	0.00169	0.02688	0.08757	0	0.03697	0.07071	0
$[2600, 2650)$	0.00285	0.03552	0.07227	0	0.0295	0.06213	0
$[2650, 2700)$	0.00441	0.04405	0.04262	0	0.05667	0.04474	0
$[2700, 2750)$	0.01005	0.05751	0.02327	0	0.18466	0.03478	0
$[2750, 2800)$	0.01385	0.0676	0.0139	0	0.19404	0.0171	0
$[2800, 2850)$	0.02272	0.0728	0.00646	0	0.20931	0.00818	0
$[2850, 2900)$	0.02643	0.09077	0.00171	0	0.16022	0.00305	0.00117
$[2900, 2950)$	0.0433	0.1158	0	0	0.08712	0	0.0021
$[2950, 3000)$	0.06893	0.12553	0	0	0.02096	0	0.00083
$[3000, 3050)$	0.08873	0.10131	0	0	0.00063	0	0
$[3050, 3100)$	0.1063	0.06574	0	0	0	0	0.00001
$[3100, 3150)$	0.12242	0.05454	0	0	0	0	0.00639
$[3150, 3200)$	0.13729	0.04528	0	0	0	0	0.02374
$[3200, 3250)$	0.13883	0.03369	0	0	0	0	0.08357
$[3250, 3300)$	0.09751	0.02581	0	0	0	0	0.13554
$[3300, 3350)$	0.05452	0.00765	0	0	0	0	0.18649
$[3350, 3400)$	0.03426	0.0021	0	0	0	0	0.26685
$[3400, 3450)$	0.01391	0.00026	0	0	0	0	0.17913
$[3450, 3500)$	0.00518	0	0	0	0	0	0.05471
$[3500, 3550)$	0.00335	0	0	0	0	0	0.01814
$[3550, 3600)$	0.00145	0	0	0	0	0	0.01043
$[3600, 3650)$	0.00095	0	0	0	0	0	0.00449
$[3650, 3700)$	0.00034	0	0	0	0	0	0.00507
$[3700, 3750)$	0	0	0	0	0	0	0.00873
$[3750, 3800)$	0	0	0	0	0	0	0.00507
$[3800, 3850)$	0	0	0	0	0	0	0.00712
$[3850, 3900)$	0	0	0	0	0	0	0.00039
							(Continued)

Y_2	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{2k}, b_{2,k+1})$	p'_{12k}	p'_{22k}	p'_{32k}	p'_{42k}	p'_{52k}	p'_{62k}	p'_{72k}
$[0, 20)$	0.09491	0.08054	0.05997	0.04186	0.03813	0.11948	0.08947
$[20, 40)$	0.09745	0.08259	0.05599	0.04951	0.04814	0.11965	0.07084
$[40, 60)$	0.09417	0.08876	0.05227	0.06953	0.0908	0.0885	0.0705
$[60, 80)$	0.08232	0.08609	0.05367	0.07499	0.13526	0.05614	0.08279
$[80, 100)$	0.06844	0.07682	0.06704	0.10193	0.13599	0.04111	0.10336
$[100, 120)$	0.05998	0.06596	0.07912	0.17801	0.09755	0.03374	0.07587
$[120, 140)$	0.0487	0.05919	0.07356	0.17546	0.06763	0.02263	0.05661
$[140, 160)$	0.03875	0.0503	0.06041	0.06225	0.05783	0.01117	0.05022
$[160, 180)$	0.03167	0.04481	0.05728	0.02912	0.05562	0.01273	0.0706
$[180, 200)$	0.02961	0.03875	0.04581	0.03386	0.03877	0.0114	0.03784
$[200, 220)$	0.02648	0.0384	0.04181	0.02039	0.03234	0.01031	0.02613
$[220, 240)$	0.02211	0.03617	0.03697	0.01857	0.0237	0.0148	0.01638
$[240, 260)$	0.0244	0.03561	0.02839	0.01602	0.02602	0.0224	0.01253
$[260, 280)$	0.03581	0.0335	0.03119	0.01383	0.03582	0.03553	0.0157
$[280, 300)$	0.04797	0.03255	0.04799	0.01129	0.02865	0.05384	0.03374
$[300, 320)$	0.06106	0.04162	0.06724	0.02657	0.02507	0.10745	0.04924
$[320, 340)$	0.06893	0.0522	0.07557	0.04077	0.0315	0.14084	0.05987
$[340, 360)$	0.06723	0.05613	0.0657	0.03604	0.03118	0.09829	0.0783

Y_3	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{3k}, b_{3,k+1})$	p'_{13k}	p'_{23k}	p'_{33k}	p'_{43k}	p'_{53k}	p'_{63k}	p'_{73k}
$[0, 5)$	0.11639	0.11125	0.03583	0.07463	0.07342	0.04008	0.08708
$[5, 10)$	0.27713	0.26495	0.10718	0.17073	0.19973	0.11556	0.23974
$[10, 15)$	0.28831	0.28129	0.16686	0.16199	0.20573	0.18771	0.31589
$[15, 20)$	0.17787	0.18366	0.18476	0.166	0.20141	0.2252	0.19488
$[20, 25)$	0.08703	0.09282	0.18504	0.17037	0.16064	0.21155	0.08406
$[25, 30)$	0.03578	0.0428	0.16555	0.13724	0.1006	0.13693	0.03788
$[30, 35)$	0.01267	0.01707	0.1027	0.08664	0.05067	0.06783	0.02116
$[35, 40)$	0.0034	0.00508	0.04274	0.02585	0.00453	0.01353	0.01365
$[40, 45)$	0.0011	0.00072	0.00828	0.00582	0.00105	0.00104	0.00405
$[45, 50)$	0.00029	0.00018	0.00106	0.00073	0.002	0.00029	0.00151
$[50, 55)$	3e-05	8e-05	0	0	0.00021	0.00029	1e-04
$[55, 60)$	0	6e-05	0	0	0	0	0
$[60, 65)$	0	4e-05	0	0	0	0	0
$[65, 70)$	0	0	0	0	0	0	0

(Continued)

Y_4	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{4k}, b_{4,k+1})$	p'_{14k}	p'_{24k}	p'_{34k}	p'_{44k}	p'_{54k}	p'_{64k}	p'_{74k}
$[0, 50)$	0.14387	0.10166	0.12555	0.57809	0.15596	0.2256	0.11177
$[50, 100)$	0.11947	0.11049	0.14762	0.06807	0.15596	0.17297	0.11177
$[100, 150)$	0.11343	0.11885	0.14684	0.06152	0.12235	0.16088	0.07989
$[150, 200)$	0.08039	0.08855	0.10374	0.04405	0.12235	0.10244	0.07989
$[200, 250)$	0.09279	0.1071	0.11943	0.05752	0.10086	0.10958	0.07111
$[250, 300)$	0.08146	0.0949	0.10474	0.05934	0.10086	0.08862	0.07111
$[300, 350)$	0.0628	0.07191	0.07842	0.04296	0.05283	0.0513	0.05283
$[350, 400)$	0.0573	0.06516	0.06108	0.03859	0.05283	0.0395	0.05283
$[400, 450)$	0.05257	0.05442	0.04581	0.02657	0.0316	0.02447	0.04807
$[450, 500)$	0.04144	0.04157	0.02959	0.0142	0.0316	0.01399	0.04807
$[500, 550)$	0.0338	0.03371	0.0177	0.00837	0.01712	0.00656	0.04174
$[550, 600)$	0.03122	0.02987	0.01225	0.00073	0.01712	0.00294	0.04174
$[600, 650)$	0.02216	0.02104	0.00498	0	0.00864	0.00115	0.02538
$[650, 700)$	0.01867	0.01645	0.0021	0	0.00864	0	0.02538
$[700, 750)$	0.01327	0.01096	0.00014	0	0.00437	0	0.01802
$[750, 800)$	0.01104	0.00893	0	0	0.00437	0	0.01802
$[800, 850)$	0.00854	0.00691	0	0	0.00321	0	0.01499
$[850, 900)$	0.00552	0.00474	0	0	0.00321	0	0.01499
$[900, 950)$	0.00389	0.00324	0	0	0.00179	0	0.0129
$[950, 1000)$	0.00322	0.00264	0	0	0.00179	0	0.0129
$[1000, 1050)$	0.00179	0.00194	0	0	0.00126	0	0.01294
$[1050, 1100)$	0.00095	0.00146	0	0	0.00126	0	0.01294
$[1100, 1150)$	0.00035	0.00127	0	0	0	0	0.008
$[1150, 1200)$	8e-05	0.00083	0	0	0	0	0.008
$[1200, 1250)$	0	0.00071	0	0	0	0	0.00229
$[1250, 1300)$	0	4e-04	0	0	0	0	0.00229
$[1300, 1350)$	0	0.00021	0	0	0	0	7e-05
$[1350, 1400)$	0	7e-05	0	0	0	0	7e-05
							(Continued)

Y_5	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{5k}, b_{5,k+1})$	p'_{15k}	p'_{25k}	p'_{35k}	p'_{45k}	p'_{55k}	p'_{65k}	p'_{75k}
$[-200, -179.76)$	0	5e-05	0	0	0	0	0
$[-179.76, -159.52)$	0	5e-05	0	0	0	0	0
$[-159.52, -139.29)$	2e-04	0.00031	1e-05	0	1e-05	1e-05	0
$[-139.29, -119.05)$	0.00038	0.00054	0.00024	0	0.00028	0.00018	0
$[-119.05, -98.81)$	5e-04	0.00069	0.0011	0	0.00164	0.00037	0.00011
$[-98.81, -78.57)$	0.00232	0.00299	0.00198	0	0.00237	0.00071	0.00187
$[-78.57, -58.33)$	0.00232	0.00299	0.00281	0	0.00429	0.00168	0.00187
$[-58.33, -38.1)$	0.0439	0.0341	0.0048	1e-04	0.00631	0.00275	0.03706
$[-38.1, -17.86)$	0.07301	0.05588	0.01847	0.04838	0.02683	0.02252	0.06169
$[-17.86, 2.38)$	0.08846	0.07363	0.09242	0.41359	0.16318	0.15985	0.07273
$[2.38, 22.62)$	0.20429	0.20677	0.17166	0.14671	0.21091	0.22101	0.15553
$[22.62, 42.86)$	0.20429	0.20677	0.16107	0.06411	0.13195	0.16372	0.15553
$[42.86, 63.1)$	0.11934	0.12751	0.13064	0.05469	0.11341	0.12421	0.10438
$[63.1, 83.33)$	0.07301	0.08428	0.11176	0.05012	0.09733	0.10354	0.07647
$[83.33, 103.57)$	0.06512	0.0748	0.08789	0.04247	0.07455	0.07812	0.07099
$[103.57, 123.81)$	0.02832	0.03055	0.06413	0.0423	0.0507	0.0525	0.04543
$[123.81, 144.05)$	0.02832	0.03055	0.04624	0.04627	0.03522	0.02882	0.04543
$[144.05, 164.29)$	0.01831	0.01854	0.03496	0.04046	0.02562	0.01713	0.03484
$[164.29, 184.52)$	0.01414	0.01353	0.025	0.02397	0.01845	0.01091	0.03043
$[184.52, 204.76)$	0.01217	0.01175	0.01776	0.01311	0.01453	0.00606	0.02732
$[204.76, 225)$	0.00578	0.00595	0.01284	0.00599	0.01168	0.0034	0.01719
$[225, 245.24)$	0.00578	0.00595	0.00771	0.00396	0.0076	0.00143	0.01719
$[245.24, 265.48)$	0.00272	0.00343	0.00424	0.00299	0.00284	0.00064	0.01085
$[265.48, 285.71)$	0.00178	0.00265	0.00177	0.00079	0.00031	0.00037	0.0089
$[285.71, 305.95)$	0.00156	0.00216	0.00044	0	0	8e-05	0.0077
$[305.95, 326.19)$	0.00104	0.00099	6e-05	0	0	0	0.00484
$[326.19, 346.43)$	0.00104	0.00099	0	0	0	0	0.00484
$[346.43, 366.67)$	0.00064	0.00041	0	0	0	0	0.00269
$[366.67, 386.9)$	0.00056	0.00029	0	0	0	0	0.00223
$[386.9, 407.14)$	0.00041	0.00021	0	0	0	0	0.00151
$[407.14, 427.38)$	0.00014	8e-05	0	0	0	0	0.00018
$[427.38, 447.62)$	0.00014	8e-05	0	0	0	0	0.00018
$[447.62, 467.86)$	2e-05	5e-05	0	0	0	0	2e-05
$[467.86, 488.1)$	0	5e-05	0	0	0	0	0
$[488.1, 508.33)$	0	6e-05	0	0	0	0	0
$[508.33, 528.57)$	0	8e-05	0	0	0	0	0
$[528.57, 548.81)$	0	8e-05	0	0	0	0	0
$[548.81, 569.05)$	0	8e-05	0	0	0	0	0
$[569.05, 589.29)$	0	8e-05	0	0	0	0	0
$[589.29, 609.52)$	0	4e-05	0	0	0	0	0
$[609.52, 629.76)$	0	0	0	0	0	0	0
$[629.76, 650)$	0	0	0	0	0	0	0
							(Continued)

Y_6	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{6k}, b_{6,k+1})$	p'_{16k}	p'_{26k}	p'_{36k}	p'_{46k}	p'_{56k}	p'_{66k}	p'_{76k}
$[0, 100)$	0.00897	0.01383	0.03615	0.00364	0.05252	0.01788	0.00015
$[100, 200)$	0.00897	0.01383	0.03615	0.0182	0.05252	0.01788	0.00015
$[200, 300)$	0.00897	0.01383	0.06566	0.03531	0.05252	0.04624	0.00015
$[300, 400)$	0.00897	0.01383	0.06566	0.05169	0.05252	0.04624	0.00015
$[400, 500)$	0.00897	0.01383	0.07325	0.06261	0.05252	0.06167	0.00015
$[500, 600)$	0.01969	0.02772	0.07325	0.06443	0.03398	0.06167	0.01428
$[600, 700)$	0.01969	0.02772	0.05942	0.06553	0.03398	0.07293	0.01428
$[700, 800)$	0.01969	0.02772	0.05942	0.07353	0.03398	0.07293	0.01428
$[800, 900)$	0.01969	0.02772	0.06173	0.08373	0.03398	0.06101	0.01428
$[900, 1000)$	0.01969	0.02772	0.06173	0.0881	0.03398	0.06101	0.01428
$[1000, 1100)$	0.02672	0.03062	0.05697	0.07754	0.02838	0.0608	0.02096
$[1100, 1200)$	0.02672	0.03062	0.05697	0.11067	0.02838	0.0608	0.02096
$[1200, 1300)$	0.02672	0.03062	0.04785	0.10339	0.02838	0.05603	0.02096
$[1300, 1400)$	0.02672	0.03062	0.04785	0.08482	0.02838	0.05603	0.02096
$[1400, 1500)$	0.02672	0.03062	0.03284	0.05461	0.02838	0.03901	0.02096
$[1500, 1600)$	0.02562	0.02614	0.03284	0.01529	0.03948	0.03901	0.02783
$[1600, 1700)$	0.02562	0.02614	0.01796	0.00655	0.03948	0.02663	0.02783
$[1700, 1800)$	0.02562	0.02614	0.01796	0.00036	0.03948	0.02663	0.02783
$[1800, 1900)$	0.02562	0.02614	0.01236	0	0.03948	0.02142	0.02783
$[1900, 2000)$	0.02562	0.02614	0.01236	0	0.03948	0.02142	0.02783
$[2000, 2100)$	0.02414	0.02084	0.01134	0	0.02857	0.01992	0.02745
$[2100, 2200)$	0.02414	0.02084	0.01134	0	0.02857	0.01992	0.02745
$[2200, 2300)$	0.02414	0.02084	0.00931	0	0.02857	0.01068	0.02745
$[2300, 2400)$	0.02414	0.02084	0.00931	0	0.02857	0.01068	0.02745
$[2400, 2500)$	0.02414	0.02084	0.00695	0	0.02857	0.00268	0.02745
$[2500, 2600)$	0.02088	0.01836	0.00695	0	0.00775	0.00268	0.02844
$[2600, 2700)$	0.02088	0.01836	0.00413	0	0.00775	0.00107	0.02844
$[2700, 2800)$	0.02088	0.01836	0.00413	0	0.00775	0.00107	0.02844
$[2800, 2900)$	0.02088	0.01836	0.00185	0	0.00775	0.00155	0.02844
$[2900, 3000)$	0.02088	0.01836	0.00185	0	0.00775	0.00155	0.02844
$[3000, 3100)$	0.01972	0.01447	0.0015	0	0.002	0.00049	0.02146
$[3100, 3200)$	0.01972	0.01447	0.0015	0	0.002	0.00049	0.02146
$[3200, 3300)$	0.01972	0.01447	7e-04	0	0.002	0	0.02146
$[3300, 3400)$	0.01972	0.01447	7e-04	0	0.002	0	0.02146
$[3400, 3500)$	0.01972	0.01447	4e-05	0	0.002	0	0.02146
$[3500, 3600)$	0.01418	0.01006	4e-05	0	0	0	0.02271
$[3600, 3700)$	0.01418	0.01006	0	0	0	0	0.02271
$[3700, 3800)$	0.01418	0.01006	0	0	0	0	0.02271

(Continued)

Y_6	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{6k}, b_{6,k+1})$	p'_{16k}	p'_{26k}	p'_{36k}	p'_{46k}	p'_{56k}	p'_{66k}	p'_{76k}
$[3800, 3900)$	0.01418	0.01006	0	0	0	0	0.02271
$[3900, 4000)$	0.01418	0.01006	0	0	0	0	0.02271
$[4000, 4100)$	0.01251	0.00931	0	0	0	0	0.02205
$[4100, 4200)$	0.01251	0.00931	0	0	0	0	0.02205
$[4200, 4300)$	0.01251	0.00931	0	0	0	0	0.02205
$[4300, 4400)$	0.01251	0.00931	0	0	0	0	0.02205
$[4400, 4500)$	0.01251	0.00931	0	0	0	0	0.02205
$[4500, 4600)$	0.00994	0.0074	0	0	0.00544	0	0.00723
$[4600, 4700)$	0.00994	0.0074	0	0	0.00544	0	0.00723
$[4700, 4800)$	0.00994	0.0074	0	0	0.00544	0	0.00723
$[4800, 4900)$	0.00994	0.0074	0	0	0.00544	0	0.00723
$[4900, 5000)$	0.00994	0.0074	0	0	0.00544	0	0.00723
$[5000, 5100)$	0.00861	0.00801	0	0	0.00188	0	0.00745
$[5100, 5200)$	0.00861	0.00801	0	0	0.00188	0	0.00745
$[5200, 5300)$	0.00861	0.00801	0	0	0.00188	0	0.00745
$[5300, 5400)$	0.00861	0.00801	0	0	0.00188	0	0.00745
$[5400, 5500)$	0.00861	0.00801	0	0	0.00188	0	0.00745
$[5500, 5600)$	0.00662	0.0089	0	0	0	0	0
$[5600, 5700)$	0.00662	0.0089	0	0	0	0	0
$[5700, 5800)$	0.00662	0.0089	0	0	0	0	0
$[5800, 5900)$	0.00662	0.0089	0	0	0	0	0
$[5900, 6000)$	0.00662	0.0089	0	0	0	0	0
$[6000, 6100)$	0.00233	0.00346	0	0	0	0	0
$[6100, 6200)$	0.00233	0.00346	0	0	0	0	0
$[6200, 6300)$	0.00233	0.00346	0	0	0	0	0
$[6300, 6400)$	0.00233	0.00346	0	0	0	0	0
$[6400, 6500)$	0.00233	0.00346	0	0	0	0	0
$[6500, 6600)$	6e-05	0.00083	0	0	0	0	0
$[6600, 6700)$	6e-05	0.00083	0	0	0	0	0
$[6700, 6800)$	6e-05	0.00083	0	0	0	0	0
$[6800, 6900)$	6e-05	0.00083	0	0	0	0	0
$[6900, 7000)$	6e-05	0.00083	0	0	0	0	0
$[7000, 7100)$	0	4e-05	0	0	0	0	0
$[7100, 7200)$	0	4e-05	0	0	0	0	0
$[7200, 7300)$	0	4e-05	0	0	0	0	0
$[7300, 7400)$	0	4e-05	0	0	0	0	0
$[7400, 7500)$	0	4e-05	0	0	0	0	0

(Continued)

Y_7	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{7k}, b_{7,k+1})$	p'_{17k}	p'_{27k}	p'_{37k}	p'_{47k}	p'_{57k}	p'_{67k}	p'_{77k}
$[0, 10)$	2e-05	2e-05	0	0	0	9e-05	0
$[10, 20)$	0	0	0	0	0	9e-05	0
$[20, 30)$	0	0	0	0	0	3e-05	0
$[30, 40)$	0	0	0	0	0	3e-05	0
$[40, 50)$	0	0	3e-05	0	0	6e-05	0
$[50, 60)$	5e-05	5e-05	0.00011	0	0	6e-05	0
$[60, 70)$	9e-05	6e-05	0.00078	0	0	0.00012	0
$[70, 80)$	0.00027	0.00018	0.00159	0	0	0.00012	0
$[80, 90)$	0.00043	0.00034	0.00372	0	0	0.00058	5e-05
$[90, 100)$	0.00065	0.00073	0.00554	0	0	0.00058	1e-04
$[100, 110)$	0.00107	0.0012	0.01189	0	0	0.00728	0.00083
$[110, 120)$	0.00222	0.00182	0.01673	0	0	0.00728	0.00122
$[120, 130)$	0.00417	0.00323	0.02363	0.00036	0.00042	0.02775	0.00141
$[130, 140)$	0.00645	0.00566	0.0337	0.00291	0.002	0.02775	0.00254
$[140, 150)$	0.00895	0.00998	0.04128	0.00364	0.00348	0.06003	0.00531
$[150, 160)$	0.01431	0.01522	0.0457	0.00837	0.00737	0.06003	0.01009
$[160, 170)$	0.02378	0.02335	0.05166	0.01638	0.01464	0.08352	0.01633
$[170, 180)$	0.04274	0.03966	0.05253	0.0233	0.0277	0.08352	0.03091
$[180, 190)$	0.06667	0.05873	0.05697	0.03058	0.04761	0.09213	0.06724
$[190, 200)$	0.09595	0.08672	0.06699	0.05315	0.06447	0.09213	0.10078
$[200, 210)$	0.13396	0.1188	0.07129	0.06516	0.07405	0.10428	0.10843
$[210, 220)$	0.17491	0.16033	0.08318	0.10885	0.11472	0.10428	0.13993
$[220, 230)$	0.18921	0.19822	0.11448	0.13578	0.18951	0.0959	0.1943
$[230, 240)$	0.14581	0.17427	0.13232	0.14161	0.19277	0.0959	0.16568
$[240, 250)$	0.07561	0.08956	0.13464	0.21478	0.20299	0.02824	0.1275
$[250, 260)$	0.01265	0.01187	0.05124	0.19512	0.05825	0.02824	0.02735
							(Continued)

Y_8	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{8k}, b_{8,k+1})$	p'_{18k}	p'_{28k}	p'_{38k}	p'_{48k}	p'_{58k}	p'_{68k}	p'_{78k}
$[0, 10)$	0	2e-05	0	0	0	0	0
$[10, 20)$	0	0	0	0	0	0	0
$[20, 30)$	0	0	0	0	0	0	0
$[30, 40)$	0	0	0	0	0	0	0
$[40, 50)$	0	1e-05	0	0	0	0	0
$[50, 60)$	0	1e-05	0	0	0	0	0
$[60, 70)$	0	1e-05	0	0	0	0	0
$[70, 80)$	1e-05	1e-05	0	0	0	0	0
$[80, 90)$	3e-05	2e-05	0	0	0	0	0
$[90, 100)$	7e-05	4e-05	0.00011	0	0.00084	6e-05	2e-04
$[100, 110)$	0.00011	6e-05	0.00039	0	0.00053	0.00012	0.00029
$[110, 120)$	0.00012	0.00012	0.00078	0	0.00126	0.00046	0.00088
$[120, 130)$	0.00028	0.00024	0.00336	0	0.00053	0.00063	0.002
$[130, 140)$	0.00058	4e-04	0.00722	0.00036	0.00053	0.00432	0.00219
$[140, 150)$	0.00133	0.00096	0.01491	0.00218	0.00158	0.01296	0.00293
$[150, 160)$	0.00278	0.00253	0.02179	0.00728	0.01338	0.02603	0.00497
$[160, 170)$	0.00576	0.00588	0.03295	0.0182	0.03424	0.03501	0.00878
$[170, 180)$	0.01282	0.01258	0.04416	0.03203	0.04245	0.04629	0.01531
$[180, 190)$	0.02866	0.02334	0.05474	0.06079	0.05594	0.07186	0.02477
$[190, 200)$	0.05254	0.04702	0.07585	0.08518	0.06426	0.10998	0.04993
$[200, 210)$	0.1	0.09411	0.10667	0.13287	0.08311	0.14919	0.12335
$[210, 220)$	0.17525	0.16101	0.1326	0.1751	0.13547	0.18293	0.19083
$[220, 230)$	0.23468	0.22437	0.14521	0.22024	0.18224	0.16065	0.21312
$[230, 240)$	0.21813	0.206	0.14706	0.1376	0.17013	0.1052	0.18679
$[240, 250)$	0.13426	0.16409	0.14605	0.09028	0.16402	0.07474	0.1511
$[250, 260)$	0.0326	0.05717	0.06615	0.03786	0.04951	0.01958	0.02257

(Continued)

Y_9	$\mathbf{y}_1'$	$\mathbf{y}_2'$	$\mathbf{y}_3'$	$\mathbf{y}_4'$	$\mathbf{y}_5'$	$\mathbf{y}_6'$	$\mathbf{y}_7'$
$[b_{9k}, b_{9,k+1})$	p_{19k}'	p_{29k}'	p_{39k}'	p_{49k}'	p_{59k}'	p_{69k}'	p_{79k}'
$[0, 10)$	0.00192	0.00166	0.00611	0.0162	0.00974	0.00014	0.0078
$[10, 20)$	0.00048	0.00071	0.00611	0.0162	0.00974	0.00014	0.0078
$[20, 30)$	0.00108	0.00151	0.01018	0.02439	0.02333	0.00423	0.00471
$[30, 40)$	0.0022	0.00259	0.01018	0.02439	0.02333	0.00423	0.00471
$[40, 50)$	0.00396	0.00374	0.02036	0.04168	0.03171	0.01111	0.0088
$[50, 60)$	0.00718	0.00624	0.02036	0.04168	0.03171	0.01111	0.0088
$[60, 70)$	0.01118	0.00987	0.03415	0.06443	0.04909	0.02519	0.02138
$[70, 80)$	0.01809	0.01631	0.03415	0.06443	0.04909	0.02519	0.02138
$[80, 90)$	0.026	0.0246	0.04471	0.06644	0.0503	0.04157	0.04054
$[90, 100)$	0.0379	0.04025	0.04471	0.06644	0.0503	0.04157	0.04054
$[100, 110)$	0.0553	0.0629	0.05997	0.06261	0.05989	0.05752	0.07301
$[110, 120)$	0.07487	0.08319	0.05997	0.06261	0.05989	0.05752	0.07301
$[120, 130)$	0.09748	0.1064	0.07107	0.06844	0.08438	0.07416	0.09476
$[130, 140)$	0.11494	0.11908	0.07107	0.06844	0.08438	0.07416	0.09476
$[140, 150)$	0.12035	0.119	0.06858	0.06862	0.07774	0.07353	0.12167
$[150, 160)$	0.10469	0.09753	0.06858	0.06862	0.07774	0.07353	0.12167
$[160, 170)$	0.08951	0.08047	0.05939	0.04951	0.05025	0.07347	0.07777
$[170, 180)$	0.07779	0.06765	0.05939	0.04951	0.05025	0.07347	0.07777
$[180, 190)$	0.06034	0.0576	0.05376	0.02439	0.04251	0.06953	0.03891
$[190, 200)$	0.04186	0.04247	0.05376	0.02439	0.04251	0.06953	0.03891
$[200, 210)$	0.02637	0.02662	0.04264	0.00965	0.01828	0.05168	0.00948
$[210, 220)$	0.01489	0.01491	0.04264	0.00965	0.01828	0.05168	0.00948
$[220, 230)$	0.00765	0.00832	0.02541	0.00364	0.00279	0.01756	0.00117
$[230, 240)$	0.00321	0.00466	0.02541	0.00364	0.00279	0.01756	0.00117
$[240, 250)$	0.00074	0.00159	0.00368	0	0	0.00029	0
$[250, 260)$	2e-05	0.00012	0.00368	0	0	0.00029	0
							(Continued)

Y_{10}	$\mathbf{y}'_1$	$\mathbf{y}'_2$	$\mathbf{y}'_3$	$\mathbf{y}'_4$	$\mathbf{y}'_5$	$\mathbf{y}'_6$	$\mathbf{y}'_7$
$[b_{10,k}, b_{10,k+1})$	$p'_{1,10,k}$	$p'_{2,10,k}$	$p'_{3,10,k}$	$p'_{4,10,k}$	$p'_{5,10,k}$	$p'_{6,10,k}$	$p'_{7,10,k}$
$[0, 197.37)$	0.02643	0.01946	0.0361	0.05856	0.02507	0.02119	0.024
$[197.37, 394.74)$	0.02643	0.01946	0.11764	0.13937	0.02507	0.07401	0.024
$[394.74, 592.11)$	0.04202	0.03611	0.16317	0.15165	0.05554	0.1157	0.03817
$[592.11, 789.47)$	0.05983	0.05513	0.15758	0.1423	0.09036	0.16521	0.05437
$[789.47, 986.84)$	0.05983	0.05513	0.14894	0.14462	0.09036	0.1554	0.05437
$[986.84, 1184.21)$	0.06688	0.07195	0.11732	0.08442	0.0883	0.11914	0.05475
$[1184.21, 1381.58)$	0.06739	0.07315	0.08932	0.10027	0.08815	0.11226	0.05477
$[1381.58, 1578.95)$	0.06703	0.07203	0.0578	0.08085	0.09126	0.07371	0.06029
$[1578.95, 1776.32)$	0.0665	0.07034	0.03935	0.06134	0.09593	0.04201	0.06855
$[1776.32, 1973.68)$	0.0665	0.07034	0.02617	0.03272	0.09593	0.03365	0.06855
$[1973.68, 2171.05)$	0.06145	0.06206	0.01383	0.00388	0.07232	0.02853	0.06078
$[2171.05, 2368.42)$	0.06067	0.06078	0.01361	0	0.06869	0.0195	0.05959
$[2368.42, 2565.79)$	0.05598	0.05475	0.01161	0	0.04958	0.01855	0.05409
$[2565.79, 2763.16)$	0.04661	0.04267	0.00573	0	0.01135	0.0145	0.04311
$[2763.16, 2960.53)$	0.04661	0.04267	0.00164	0	0.01135	0.00583	0.04311
$[2960.53, 3157.89)$	0.03024	0.02097	0.00019	0	0.0036	0.00081	0.03791
$[3157.89, 3355.26)$	0.02615	0.01554	0	0	0.00166	0	0.03661
$[3355.26, 3552.63)$	0.02298	0.01374	0	0	0.00145	0	0.03682
$[3552.63, 3750)$	0.01426	0.00881	0	0	0.00087	0	0.0374
$[3750, 3947.37)$	0.01426	0.00881	0	0	0.00087	0	0.0374
$[3947.37, 4144.74)$	0.00965	0.00965	0	0	0.00023	0	0.02183
$[4144.74, 4342.11)$	0.00798	0.00995	0	0	0	0	0.01617
$[4342.11, 4539.47)$	0.00752	0.01003	0	0	0	0	0.01297
$[4539.47, 4736.84)$	0.0057	0.01035	0	0	0	0	0.00017
$[4736.84, 4934.21)$	0.0057	0.01035	0	0	0	0	0.00017
$[4934.21, 5131.58)$	0.00519	0.0094	0	0	0.00208	0	6e-05
$[5131.58, 5328.95)$	0.00493	0.00892	0	0	0.00312	0	0
$[5328.95, 5526.32)$	0.00476	0.00897	0	0	0.0032	0	0
$[5526.32, 5723.68)$	0.00366	0.00924	0	0	0.0037	0	0
$[5723.68, 5921.05)$	0.00366	0.00924	0	0	0.0037	0	0
$[5921.05, 6118.42)$	0.00332	0.00885	0	0	0.00497	0	0
$[6118.42, 6315.79)$	0.00309	0.00859	0	0	0.00582	0	0
$[6315.79, 6513.16)$	0.00298	0.00813	0	0	0.00543	0	0
$[6513.16, 6710.53)$	0.00147	0.00171	0	0	0	0	0
$[6710.53, 6907.89)$	0.00147	0.00171	0	0	0	0	0
$[6907.89, 7105.26)$	0.00073	0.00084	0	0	0	0	0
$[7105.26, 7302.63)$	8e-05	8e-05	0	0	0	0	0
$[7302.63, 7500)$	8e-05	8e-05	0	0	0	0	0

B.3 Dissimilarity Matrices

In order to obtain dissimilarity matrices, first of all, the histogram-valued data for the cover type dataset shown in Table B.1 should be transformed using the method proposed in Section 3.2. These transformed histogram-valued data for seven objects with ten variables are shown in Table B.2. By this transformation, observations have the same number and length of subintervals for each variable as shown in Table B.2. We can calculate dissimilarity matrices for the transformed histogram-valued cover type data. As mentioned in Section 5.2, for the cluster analysis, we use four different dissimilarity measures, the extended Gowda-Diday measure (GD), the city block (CB) and Euclidean distances (EU) (based on the normalized extended Ichino-Yaguchi measure for $\gamma = 0.25$), and the normalized CDF measure (NCDF).

Using Equation (3.8), we can calculate the extended Gowda-Diday dissimilarity measure as follows: we first have to compute the intersection of $\mathbf{y}'_{i_1} = \mathbf{y}'_1$ and $\mathbf{y}'_{i_2} = \mathbf{y}'_2$ using the transformed histogram-valued data of Table B.2. For the variable $Y_j = Y_1$,

$$p'_{(i_1 \cap i_2)jk} = p'_{(1\cap 2)11} = \min\{0, 0\} = 0.$$

where subscripts i_1 and i_2 are the i_1^{th} and i_2^{th} observations, respectively, j is the j^{th} variable, and k is the k^{th} subinterval or relative frequency. Similarly,

$$\begin{aligned}
&p'_{(1\cap 2)12} = \min\{0, 0\} = 0, \ldots, \ p'_{(1\cap 2)1,13} = \min\{0, 0.00522\} = 0, \\
&p'_{(1\cap 2)1,14} = \min\{0.0007, 0.01775\} = 0.0007, \\
&p'_{(1\cap 2)1,15} = \min\{0.00169, 0.02688\} = 0.00169, \ldots, \\
&p'_{(1\cap 2)1,32} = \min\{0.01391, 0.00026\} = 0.00026, \\
&p'_{(1\cap 2)1,33} = \min\{0.00518, 0\} = 0, \ldots, \\
&p'_{(1\cap 2)1,41} = \min\{0, 0\} = 0.
\end{aligned}$$

Thus, for $i_1 = 1,\ i_2 = 2,\ j = 1$, $y'_{(i_1 \cap i_2)j} = y'_{(1\cap 2)1}$ is

$$\begin{aligned} y'_{(1\cap 2)1} \quad = \quad & \big\{[1850, 1900), 0; \cdots ;\ [2450, 2500), 0;\ [2500, 2550), 0.0007; \\ & [2550, 2600), 0.00169; \cdots ;\ [3400, 3450), 0.00026; \\ & [3450, 3500), 0; \cdots ;\ [3850, 3900), 0\big\}. \end{aligned}$$

Now, we calculate the mean and standard deviation of $y'_{i_1 j} = y'_{11}$. These are, respectively,

$$\begin{aligned} M_{i_1 j} = M_{11} \quad = \quad & \frac{1}{2}\big\{0 + \cdots + 0 + (2500 + 2550)0.0007 + (2550 + 2600)0.00169 + \cdots \\ & + (3600 + 3650)0.00095 + ((3650 + 3700)0.0034 + 0 + \cdots + 0\big\} \\ = \quad & 3128.218, \end{aligned}$$

and

$$\begin{aligned} S_{i_1 j} \quad = \quad & S_{11} \\ = \quad & \Bigg\{0 + \cdots + \left(\frac{(-625.218)^2 + (-625.218)(-575.218) + (-575.218)^2}{3}\right) 0.0007 + \cdots \\ & + \left(\frac{(274.782)^2 + (274.782)(324.782) + (324.782)^2}{3}\right) 0.00026 + 0 + \cdots + 0\Bigg\}^{1/2} \\ = \quad & 159.140. \end{aligned}$$

Similarly, $M_{i_2 j} = M_{21} = 2920.412$, $S_{i_2 j} = S_{21} = 187.599$.

In order to compute $S_{(i_1\cap i_2)j} = S_{(1\cap 2)1}$, first of all, we obtain $p^*_{(i_1\cap i_2)jk} = p^*_{(1\cap 2)1k},\ k = 1, \ldots, 41$, and then calculate $M^*_{(i_1\cap i_2)j} = M^*_{(1\cap 2)1}$ using Equation (3.17). Thus,

$$\sum_{k=1}^{41} p'_{(1\cap 2)1k} = 0 + \cdots + 0.0007 + 0.00169 + \cdots + 0.00026 + 0 + \cdots + 0 = 0.519.$$

Hence, from Equation (3.18), the standardized relative frequencies $p^*_{(i_1\cap i_2)jk} = p^*_{(1\cap 2)1k},\ k = 1, \ldots, 41$, are

$$\begin{aligned} & p^*_{(1\cap 2)11} = \frac{0}{0.519} = 0, \ldots, p^*_{(1\cap 2)1,13} = 0,\ p^*_{(1\cap 2)14} = \frac{0.0007}{0.519} = 0.001, \\ & p^*_{(1\cap 2)1,15} = \frac{0.00169}{0.519} = 0.003, \ldots,\ p^*_{(1\cap 2)1,32} = \frac{0.00026}{0.519} = 0.001, \\ & p^*_{(1\cap 2)1,33} = \frac{0}{0.519} = 0, \ldots,\ p^*_{(1\cap 2)1,41} = 0, \end{aligned}$$

and

$$
\begin{aligned}
M^*_{(i_1\cap i_2)j} = M^*_{(1\cap 2)1} &= \frac{1}{2}\big\{0+\cdots+(2500+2550)0.001+(2550+2600)0.003+\cdots \\
&\qquad +(3400+3450)0.001+\cdots+0\big\} = 3035.174.
\end{aligned}
$$

Hence, the standard deviation for the intersection of $y'_{i_1j} = y'_{11}$ and $y'_{i_2j} = y'_{21}$ is, from Equation (3.20),

$$
\begin{aligned}
S_{(i_1\cap i_2)j} &= S_{(1\cap 2)1} \\
&= \left\{0+\cdots+\left(\frac{(-535.174)^2+(-535.174)(-485.174)+(-485.174)^2}{3}\right)0.0007\right. \\
&\qquad \left. +\left(\frac{(364.826)^2+(364.826)(414.826)+(414.826)^2}{3}\right)0.00026+\cdots+0\right\}^{1/2} \\
&= 105.528.
\end{aligned}
$$

By using these mean and standard deviation values, the three components of the extended Gowda-Diday measure between observations $\mathbf{y}'_1$ and $\mathbf{y}'_2$ for variable Y_1 can be obtained as follows:

$$
\begin{aligned}
D_{11}(y'_{11}, y'_{21}) &= \frac{|159.140 - 187.599|}{159.140 + 187.599} = 0.082, \\
D_{21}(y'_{11}, y'_{21}) &= \frac{159.140 + 187.599 - 2(105.528)}{159.140 + 187.599} = 0.391, \\
D_{31}(y'_{11}, y'_{21}) &= \frac{|3128.218 - 2920.412|}{3900 - 1850} = 0.101.
\end{aligned}
$$

Thus,

$$
D_{GD}(y'_{11}, y'_{21}) = 0.082 + 0.391 + 0.101 = 0.575.
$$

Similarly, the extended Gowda-Diday measure values between $y'_{i_1j} = y'_{11}$ and $y'_{i_2j} = y'_{21}$ for variables $Y_2, \ldots, Y_{10}$, $D_{GD}(y'_{1j}, y'_{2j}),\ j = 2, \ldots, 10$, are

$$
\begin{aligned}
&D_{GD}(y'_{12}, y'_{22}) = 0.105,\ D_{GD}(y'_{13}, y'_{23}) = 0.056,\ D_{GD}(y'_{14}, y'_{24}) = 0.073, \\
&D_{GD}(y'_{15}, y'_{25}) = 0.053,\ D_{GD}(y'_{16}, y'_{26}) = 0.138,\ D_{GD}(y'_{17}, y'_{27}) = 0.040, \\
&D_{GD}(y'_{18}, y'_{28}) = 0.063,\ D_{GD}(y'_{19}, y'_{29}) = 0.030,\ D_{GD}(y'_{1,10}, y'_{2,10}) = 0.208.
\end{aligned}
$$

Thus, the extended Gowda-Diday dissimilarity measure between $\mathbf{y}'_1$ and $\mathbf{y}'_2$ is given by, from Equation (3.21),

$$D_{GD}(\mathbf{y}'_1, \mathbf{y}'_2) = 0.575 + 0.105 + 0.056 + \cdots + 0.208 = 1.341.$$

Similarly, we can calculate these dissimilarities $D_{GD}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2})$ for all $i_1, i_2 = 1, \ldots, 7$. Hence, we can complete the dissimilarity matrix. The extended Gowda-Diday dissimilarity matrix for the data of Table B.2 is given in Equation (5.1).

Now, we calculate the normalized city block and Euclidean distances. To obtain these distances, we first calculate the extended Ichino-Yaguchi measure of Equation (3.22). To obtain the extended Ichino-Yaguchi measure, we first compute the union between two transformed histogram-valued objects $\mathbf{y}'_{i_1} = \mathbf{y}'_1$ and $\mathbf{y}'_{i_2} = \mathbf{y}_2$. For the variable $Y_j = Y_1$,

$$p'_{(i_1 \cup i_2)jk} = p'_{(1 \cup 2)11} = \max\{0, 0\} = 0.$$

Similarly,

$$\begin{aligned}
&p'_{(1\cup 2)12} = \max\{0, 0\} = 0, \ldots,\ p'_{(1\cup 2)17} = \max\{0, 0.00015\} = 0.00015,\\
&p'_{(1\cup 2)18} = \max\{0, 0.00025\} = 0.00025, \ldots,\ p'_{(1\cup 2)1,37} = \max\{0.00034, 0\} = 0.00034,\\
&p'_{(1\cup 2)1,38} = \max\{0, 0\} = 0, \ldots,\ p'_{(1\cup 2)1,41} = \max\{0, 0\} = 0.
\end{aligned}$$

Thus, $y'_{(i_1 \cup i_2)j} = y'_{(1\cup 2)1}$ is

$$\begin{aligned}
y'_{(1\cup 2)1} \;=\; &\big\{[1850, 1900), 0; \cdots;\ [2150, 2200), 0.00015;\ [2200, 2250), 0.00025; \cdots\\
&[3650, 3700), 0.00034;\ [3700, 3750), 0; \cdots;\ [3850, 3900), 0\big\}.
\end{aligned}$$

In order to compute $S_{(i_1 \cup i_2)j} = S_{(1\cup 2)1}$, first of all, we obtain $p^*_{(1_1 \cup i_2)jk} = p^*_{(1\cup 2)1k}$, $k = 1, \ldots, 41$, and then calculate $M^*_{(i_1 \cup i_2)j} = M^*_{(1\cup 2)1}$ using Equation (3.15). Hence,

$$\sum_{k=1}^{41} p'_{(1\cup 2)1k} = 0 + \cdots + 0 + 0.00015 + \cdots + 0.00034 + 0 + \cdots + 0 = 1.481.$$

Therefore, from Equation (3.16), the standardized relative frequencies $p^*_{(i_1\cup i_2)jk} = p^*_{(1\cup 2)1k}$, $k = 1, \ldots, 41$, are

$$\begin{aligned}
&p^*_{(1\cup 2)11} = 0,\ p^*_{(1\cup 2)12} = 0, \ldots,\ p^*_{(1\cup 2)17} = \frac{0.00015}{1.481} = 0.0001,\\
&p^*_{(1\cup 2)18} = \frac{0.00025}{1.481} = 0.0002, \ldots,\ p^*_{(1\cup 2)1,37} = \frac{0.00034}{1.481} = 0.0002,\\
&p^*_{(1\cup 2)1,38} = 0, \ldots,\ p^*_{(1\cup 2)1,41} = 0,
\end{aligned}$$

and

$$\begin{aligned}
M^*_{(i_1\cup i_2)j} &= M^*_{(1\cup 2)1}\\
&= \frac{1}{2}\big[0 + \cdots + (2150 + 2200)0.00015 + \cdots + (3650 + 3700)0.00034 + \cdots + 0\big]\\
&= 3020.512.
\end{aligned}$$

Thus, the standard deviation for the union of $y'_{i_1j} = y'_{11}$ and $y'_{i_2j} = y'_{21}$ is, from Equation (3.19),

$$\begin{aligned}
S_{(i_1\cup i_2)j} &= S_{(1\cup 2)1}\\
&= \Bigg\{0 + \cdots + \left(\frac{(-870.512)^2 + (-870.512)(-820.512) + (-820.512)^2}{3}\right)0.0001\\
&\qquad + \cdots + \left(\frac{(629.488)^2 + (629.488)(679.488) + (679.488)^2}{3}\right)0.0002 + \cdots + 0\Bigg\}^{1/2}\\
&= 266.255.
\end{aligned}$$

From the calculation of the extended Gowda-Diday measure, we know $S_{i_1j} = S_{11} = 159.140$, $S_{i_2j} = S_{21} = 187.599$, and $S_{(i_1\cap i_2)j} = S_{(1\cap 2)1} = 105.528$. Thus, since we assume $\gamma = 0.25$, the extended Ichino-Yaguchi measure $\phi(y'_{i_1j}, y'_{i_2j}) = \phi(y'_{11}, y'_{21})$ is, from Equation (3.22),

$$\phi(y'_{11}, y'_{21}) = 266.255 - 105.528 + 0.25(2 \times 105.528 - 159.140 - 187.599) = 126.806.$$

Similarly, for variables $Y_2, \ldots, Y_{10}$, respectively,

$$\begin{aligned}
&\phi(y'_{12}, y'_{22}) = 8.815,\ \phi(y'_{13}, y'_{23}) = 0.344,\ \phi(y'_{14}, y'_{24}) = 16.411,\\
&\phi(y'_{15}, y'_{25}) = 4.231,\ \phi(y'_{16}, y'_{26}) = 176.315,\ \phi(y'_{17}, y'_{27}) = 1.169,\\
&\phi(y'_{18}, y'_{28}) = 1.291,\ \phi(y'_{19}, y'_{29}) = 1.299,\ \phi(y'_{1,10}, y'_{2,10}) = 225.443.
\end{aligned}$$

Now we compute the normalized extended Ichino-Yaguchi measure for $\phi(y'_{11}, y'_{21})$. First, from Equation (3.24),

$$\begin{aligned}
A_{1j} = A_{11} &= 1850^2 + 1900^2 + 3850^2 + 3900^2 = 37065000, \\
A_{2j} = A_{21} &= 1850 \times 1900 + 3850 \times 3900 = 18530000, \\
A_{3j} = A_{31} &= 1850 \times 3850 + 1850 \times 3900 + 1900 \times 3850 + 1900 \times 3900 = 29062500.
\end{aligned}$$

Thus,

$$V_j = V_1 = \sqrt{\frac{5 \times 37065000 + 2 \times 18530000 - 6 \times 29062500}{24}} = 1414.361.$$

Therefore, substituting into Equation (3.23), we obtain

$$\phi^*(y'_{11}, y'_{21}) = \frac{126.806}{1414.361} = 0.090.$$

Similarly, for variables $Y_2, \ldots, Y_{10}$, respectively,

$$\begin{aligned}
&\phi^*(y'_{12}, y'_{22}) = 0.037,\ \phi^*(y'_{13}, y'_{23}) = 0.007,\ \phi^*(y'_{14}, y'_{24}) = 0.017, \\
&\phi^*(y'_{15}, y'_{25}) = 0.007,\ \phi^*(y'_{16}, y'_{26}) = 0.034,\ \phi^*(y'_{17}, y'_{27}) = 0.007, \\
&\phi^*(y'_{18}, y'_{28}) = 0.007,\ \phi^*(y'_{19}, y'_{29}) = 0.007,\ \phi^*(y'_{1,10}, y'_{2,10}) = 0.044.
\end{aligned}$$

The complete results for the normalized extended Ichino-Yaguchi dissimilarity matrices for each variable (i.e., $Y_j,\ j = 1, \ldots, 10$), ϕ^*_j, are as follows:

$$\phi^*_1 = \begin{pmatrix}
0 & 0.090 & 0.340 & 0.425 & 0.154 & 0.328 & 0.109 \\
0.090 & 0 & 0.242 & 0.325 & 0.067 & 0.230 & 0.205 \\
0.340 & 0.242 & 0 & 0.084 & 0.184 & 0.017 & 0.455 \\
0.425 & 0.325 & 0.084 & 0 & 0.263 & 0.096 & 0.541 \\
0.154 & 0.067 & 0.184 & 0.263 & 0 & 0.171 & 0.268 \\
0.328 & 0.230 & 0.017 & 0.096 & 0.171 & 0 & 0.443 \\
0.109 & 0.205 & 0.455 & 0.541 & 0.268 & 0.443 & 0
\end{pmatrix},$$

$$\phi_2^* = \begin{pmatrix} 0 & 0.037 & 0.043 & 0.102 & 0.098 & 0.072 & 0.036 \\ 0.037 & 0 & 0.054 & 0.072 & 0.067 & 0.105 & 0.038 \\ 0.043 & 0.054 & 0 & 0.091 & 0.097 & 0.099 & 0.052 \\ 0.102 & 0.072 & 0.091 & 0 & 0.041 & 0.174 & 0.080 \\ 0.098 & 0.067 & 0.097 & 0.041 & 0 & 0.173 & 0.094 \\ 0.072 & 0.105 & 0.099 & 0.174 & 0.173 & 0 & 0.095 \\ 0.036 & 0.038 & 0.052 & 0.080 & 0.094 & 0.095 & 0 \end{pmatrix},$$

$$\phi_3^* = \begin{pmatrix} 0 & 0.007 & 0.099 & 0.071 & 0.045 & 0.073 & 0.021 \\ 0.007 & 0 & 0.092 & 0.065 & 0.039 & 0.066 & 0.018 \\ 0.099 & 0.092 & 0 & 0.028 & 0.054 & 0.025 & 0.077 \\ 0.071 & 0.065 & 0.028 & 0 & 0.028 & 0.027 & 0.051 \\ 0.045 & 0.039 & 0.054 & 0.028 & 0 & 0.030 & 0.037 \\ 0.073 & 0.066 & 0.025 & 0.027 & 0.030 & 0 & 0.057 \\ 0.021 & 0.018 & 0.077 & 0.051 & 0.037 & 0.057 & 0 \end{pmatrix},$$

$$\phi_4^* = \begin{pmatrix} 0 & 0.017 & 0.065 & 0.120 & 0.039 & 0.094 & 0.076 \\ 0.017 & 0 & 0.066 & 0.124 & 0.043 & 0.099 & 0.074 \\ 0.065 & 0.066 & 0 & 0.058 & 0.035 & 0.033 & 0.138 \\ 0.120 & 0.124 & 0.058 & 0 & 0.078 & 0.024 & 0.200 \\ 0.039 & 0.043 & 0.035 & 0.078 & 0 & 0.055 & 0.115 \\ 0.094 & 0.099 & 0.033 & 0.024 & 0.055 & 0 & 0.173 \\ 0.076 & 0.074 & 0.138 & 0.200 & 0.115 & 0.173 & 0 \end{pmatrix},$$

$$\phi_5^* = \begin{pmatrix} 0 & 0.007 & 0.025 & 0.029 & 0.022 & 0.025 & 0.035 \\ 0.007 & 0 & 0.024 & 0.030 & 0.021 & 0.025 & 0.035 \\ 0.025 & 0.024 & 0 & 0.025 & 0.012 & 0.023 & 0.035 \\ 0.029 & 0.030 & 0.025 & 0 & 0.020 & 0.026 & 0.050 \\ 0.022 & 0.021 & 0.012 & 0.020 & 0 & 0.014 & 0.043 \\ 0.025 & 0.025 & 0.023 & 0.026 & 0.014 & 0 & 0.055 \\ 0.035 & 0.035 & 0.035 & 0.050 & 0.043 & 0.055 & 0 \end{pmatrix},$$

$$\phi_6^* = \begin{pmatrix} 0 & 0.034 & 0.226 & 0.252 & 0.152 & 0.219 & 0.061 \\ 0.034 & 0 & 0.217 & 0.245 & 0.145 & 0.210 & 0.088 \\ 0.226 & 0.217 & 0 & 0.043 & 0.074 & 0.020 & 0.227 \\ 0.252 & 0.245 & 0.043 & 0 & 0.112 & 0.035 & 0.250 \\ 0.152 & 0.145 & 0.074 & 0.112 & 0 & 0.077 & 0.153 \\ 0.219 & 0.210 & 0.020 & 0.035 & 0.077 & 0 & 0.219 \\ 0.061 & 0.088 & 0.227 & 0.250 & 0.153 & 0.219 & 0 \end{pmatrix},$$

$$\phi_7^* = \begin{pmatrix} 0 & 0.007 & 0.074 & 0.056 & 0.041 & 0.071 & 0.021 \\ 0.007 & 0 & 0.076 & 0.052 & 0.036 & 0.076 & 0.019 \\ 0.074 & 0.076 & 0 & 0.109 & 0.094 & 0.046 & 0.082 \\ 0.056 & 0.052 & 0.109 & 0 & 0.018 & 0.125 & 0.036 \\ 0.041 & 0.036 & 0.094 & 0.018 & 0 & 0.109 & 0.020 \\ 0.071 & 0.076 & 0.046 & 0.125 & 0.109 & 0 & 0.087 \\ 0.021 & 0.019 & 0.082 & 0.036 & 0.020 & 0.087 & 0 \end{pmatrix},$$

$$\phi_8^* = \begin{pmatrix} 0 & 0.007 & 0.046 & 0.024 & 0.032 & 0.050 & 0.013 \\ 0.007 & 0 & 0.045 & 0.030 & 0.031 & 0.057 & 0.016 \\ 0.046 & 0.045 & 0 & 0.032 & 0.020 & 0.023 & 0.038 \\ 0.024 & 0.030 & 0.032 & 0 & 0.022 & 0.028 & 0.026 \\ 0.032 & 0.031 & 0.020 & 0.022 & 0 & 0.031 & 0.030 \\ 0.050 & 0.057 & 0.023 & 0.028 & 0.031 & 0 & 0.043 \\ 0.013 & 0.016 & 0.038 & 0.026 & 0.030 & 0.043 & 0 \end{pmatrix},$$

$$\phi_9^* = \begin{pmatrix} 0 & 0.007 & 0.078 & 0.114 & 0.085 & 0.051 & 0.041 \\ 0.007 & 0 & 0.076 & 0.113 & 0.085 & 0.052 & 0.043 \\ 0.078 & 0.076 & 0 & 0.099 & 0.068 & 0.044 & 0.070 \\ 0.114 & 0.113 & 0.099 & 0 & 0.032 & 0.127 & 0.078 \\ 0.085 & 0.085 & 0.068 & 0.032 & 0 & 0.095 & 0.059 \\ 0.051 & 0.052 & 0.044 & 0.127 & 0.095 & 0 & 0.061 \\ 0.041 & 0.043 & 0.070 & 0.078 & 0.059 & 0.061 & 0 \end{pmatrix},$$

$$\phi_{10}^* = \begin{pmatrix} 0 & 0.044 & 0.159 & 0.172 & 0.064 & 0.141 & 0.058 \\ 0.044 & 0 & 0.194 & 0.206 & 0.081 & 0.175 & 0.090 \\ 0.159 & 0.194 & 0 & 0.021 & 0.106 & 0.019 & 0.150 \\ 0.172 & 0.206 & 0.021 & 0 & 0.118 & 0.034 & 0.162 \\ 0.064 & 0.081 & 0.106 & 0.118 & 0 & 0.090 & 0.106 \\ 0.141 & 0.175 & 0.019 & 0.034 & 0.090 & 0 & 0.130 \\ 0.058 & 0.090 & 0.150 & 0.162 & 0.106 & 0.130 & 0 \end{pmatrix}.$$

By using $\phi^*(y'_{1j}, y'_{2j}),\ j = 1, \ldots, 10$, the normalized city block distance of $\mathbf{y}'_1$ and $\mathbf{y}'_2$ can be computed, from Equation (3.28) with $q = 1$, as follows:

$$D^1_{NM}(\mathbf{y}_1, \mathbf{y}_2) = \big[0.090 + 0.037 + 0.007 + \cdots + 0.007 + 0.044\big] = 0.257.$$

Similarly, the normalized city block distances for all $i_1, i_2 = 1, \ldots, 7$ can be calculated. Hence, the normalized city block distance matrix is given in Equation (5.2).

From the normalized extended Ichino-Yaguchi measure values, $\phi^*(y'_{1j}, y'_{2j}),\ j = 1, \ldots, 10$, the normalized Euclidean distance between $\mathbf{y}'_1$ and $\mathbf{y}'_2$ can be calculated, from Equation (3.28) with $q = 2$, as follows:

$$D^2_{NM}(\mathbf{y}_1, \mathbf{y}_2) = \left[0.090^2 + 0.037^2 + 0.007^2 + \cdots + 0.007^2 + 0.044^2\right]^{1/2} = 0.114.$$

Similarly, the normalized Euclidean distances for all $i_1, i_2 = 1, \ldots, 7$ can be calculated. Hence, the normalized Euclidean distance matrix is given in Equation (5.3).

From the transformed histogram-valued data of Table B.2, we calculate the normalized CDF dissimilarity measure for seven cover types. First of all, the cumulative relative frequencies for $y'_{i_1j} = y'_{11}$ are obtained by

$$F_{i_1jk} = F_{111} = p'_{111} = 0,$$

Similarly,

$$\begin{aligned}
&F_{112} = F_{111} + p'_{112} = 0 + 0 = 0, \ldots,\\
&F_{11,13} = F_{11,12} + p'_{11,13} = 0 + 0 = 0,\\
&F_{11,14} = F_{11,13} + p'_{11,14} = 0 + 0.0007 = 0.0007,\\
&F_{11,15} = F_{11,14} + p'_{11,15} = 0.0007 + 0.00169 = 0.00239, \ldots,\\
&F_{11,37} = F_{11,36} + p'_{11,37} = 0.99966 + 0.00034 = 1, \ldots,\\
&F_{11,41} = F_{11,41} + p'_{11,40} = 1 + 0 = 1.
\end{aligned}$$

Similarly, cumulative relative frequencies for $y'_{i_2j} = y'_{21}$ are

$$\begin{aligned}
&F_{211} = 0, \ldots,\ F_{21,13} = 0.00942,\ F_{21,14} = 0.02717,\\
&F_{21,15} = 0.05405, \ldots,\ F_{21,37} = 1, \ldots,\ F_{21,41} = 1.
\end{aligned}$$

Thus, since $T_j = T_1 = b_{1,k+1} - b_{1k} = 50$ for $k = 1, \ldots, t_1$ from Equation (3.30), the CDF dissimilarity measure of y'_{11} and y'_{21} is obtained by

$$\begin{aligned} D_{CDF}(y'_{11}, y'_{21}) &= 50\Big[0 + \cdots + \big|0 - 0.00942\big| + \big|0.0007 - 0.02717\big| \\ &\qquad + \cdots + \big|1 - 1\big|\Big] = 207.805. \end{aligned}$$

Since $\Psi_j = \Psi_1 = b_{1,t_1+1} - b_{11} = 3900 - 1850 = 2050$, the normalized CDF measure of y'_{11} and y'_{21} is calculated as follows:

$$D_{NCDF}(y'_{11}, y'_{21}) = \frac{D_{CDF}(y'_{11}, y'_{21})}{\Psi_1} = \frac{207.805}{2050} = 0.101.$$

Similarly, for variables $Y_2, \ldots, Y_{10}$, respectively,

$$\begin{aligned} &D_{NCDF}(y'_{12}, y'_{22}) = 0.029,\ D_{NCDF}(y'_{13}, y'_{23}) = 0.006,\ D_{NCDF}(y'_{14}, y'_{24}) = 0.010, \\ &D_{NCDF}(y'_{15}, y'_{25}) = 0.005,\ D_{NCDF}(y'_{16}, y'_{26}) = 0.032,\ D_{NCDF}(y'_{17}, y'_{27}) = 0.007, \\ &D_{NCDF}(y'_{18}, y'_{28}) = 0.007,\ D_{NCDF}(y'_{19}, y'_{29}) = 0.006,\ D_{NCDF}(y'_{1,10}, y'_{2,10}) = 0.021. \end{aligned}$$

Thus, from Equation (3.31), the normalized CDF dissimilarity measure of $\mathbf{y}'_1$ and $\mathbf{y}'_2$ is

$$D_{NCDF}(\mathbf{y}'_1, \mathbf{y}'_2) = 0.101 + 0.029 + 0.006 + \cdots + 0.006 + 0.021 = 1.112.$$

Similarly, we can calculate these distances $D_{NCDF}(\mathbf{y}'_{i_1}, \mathbf{y}'_{i_2})$ for all $i_1, i_2 = 1, \ldots, 7$. Hence, we can complete the normalized CDF dissimilarity measure for the histogram-valued data of Table B.2, and this dissimilarity matrix is given in Equation (5.4).

B.4 Cluster Analysis

Table B.3: Objects sorted by mean values for each variable.

Y_1	Sort	$\mathbf{y}^1_{(1)}$	$\mathbf{y}^1_{(2)}$	$\mathbf{y}^1_{(3)}$	$\mathbf{y}^1_{(4)}$	$\mathbf{y}^1_{(5)}$	$\mathbf{y}^1_{(6)}$	$\mathbf{y}^1_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_3$	$\mathbf{y}_6$	$\mathbf{y}_5$	$\mathbf{y}_2$	$\mathbf{y}_1$	$\mathbf{y}_7$
	Mean	2223.6	2394.2	2418.8	2786.8	2920.4	3128.2	3361.7
Y_2	Sort	$\mathbf{y}^2_{(1)}$	$\mathbf{y}^2_{(2)}$	$\mathbf{y}^2_{(3)}$	$\mathbf{y}^2_{(4)}$	$\mathbf{y}^2_{(5)}$	$\mathbf{y}^2_{(6)}$	$\mathbf{y}^2_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_2$	$\mathbf{y}_7$	$\mathbf{y}_1$	$\mathbf{y}_3$	$\mathbf{y}_6$
	Mean	136.8	138.9	151.7	152.9	155.8	176.0	180.1
Y_3	Sort	$\mathbf{y}^3_{(1)}$	$\mathbf{y}^3_{(2)}$	$\mathbf{y}^3_{(3)}$	$\mathbf{y}^3_{(4)}$	$\mathbf{y}^3_{(5)}$	$\mathbf{y}^3_{(6)}$	$\mathbf{y}^3_{(7)}$
	Object	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_7$	$\mathbf{y}_5$	$\mathbf{y}_4$	$\mathbf{y}_6$	$\mathbf{y}_3$
	Mean	12.6	13.1	13.8	16.1	18.1	18.6	20.3
Y_4	Sort	$\mathbf{y}^4_{(1)}$	$\mathbf{y}^4_{(2)}$	$\mathbf{y}^4_{(3)}$	$\mathbf{y}^4_{(4)}$	$\mathbf{y}^4_{(5)}$	$\mathbf{y}^4_{(6)}$	$\mathbf{y}^4_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_6$	$\mathbf{y}_3$	$\mathbf{y}_5$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_7$
	Mean	115.5	160.1	209.3	216.3	270.3	279.2	358.4
Y_5	Sort	$\mathbf{y}^5_{(1)}$	$\mathbf{y}^5_{(2)}$	$\mathbf{y}^5_{(3)}$	$\mathbf{y}^5_{(4)}$	$\mathbf{y}^5_{(5)}$	$\mathbf{y}^5_{(6)}$	$\mathbf{y}^5_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_1$	$\mathbf{y}_6$	$\mathbf{y}_2$	$\mathbf{y}_5$	$\mathbf{y}_3$	$\mathbf{y}_7$
	Mean	37.1	42.0	44.2	46.1	49.6	61.7	68.6
Y_6	Sort	$\mathbf{y}^6_{(1)}$	$\mathbf{y}^6_{(2)}$	$\mathbf{y}^6_{(3)}$	$\mathbf{y}^6_{(4)}$	$\mathbf{y}^6_{(5)}$	$\mathbf{y}^6_{(6)}$	$\mathbf{y}^6_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_3$	$\mathbf{y}_6$	$\mathbf{y}_5$	$\mathbf{y}_2$	$\mathbf{y}_1$	$\mathbf{y}_7$
	Mean	912.8	942.9	1037.3	1354.8	2429.6	2614.4	2738.4
Y_7	Sort	$\mathbf{y}^7_{(1)}$	$\mathbf{y}^7_{(2)}$	$\mathbf{y}^7_{(3)}$	$\mathbf{y}^7_{(4)}$	$\mathbf{y}^7_{(5)}$	$\mathbf{y}^7_{(6)}$	$\mathbf{y}^7_{(7)}$
	Object	$\mathbf{y}_6$	$\mathbf{y}_3$	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_7$	$\mathbf{y}_5$	$\mathbf{y}_4$
	Mean	192.4	201.5	211.5	213.3	216.5	223.0	228.0
Y_8	Sort	$\mathbf{y}^8_{(1)}$	$\mathbf{y}^8_{(2)}$	$\mathbf{y}^8_{(3)}$	$\mathbf{y}^8_{(4)}$	$\mathbf{y}^8_{(5)}$	$\mathbf{y}^8_{(6)}$	$\mathbf{y}^8_{(7)}$
	Object	$\mathbf{y}_6$	$\mathbf{y}_3$	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_7$	$\mathbf{y}_1$	$\mathbf{y}_2$
	Mean	209.3	215.4	216.5	218.6	221.3	223.0	224.9
Y_9	Sort	$\mathbf{y}^9_{(1)}$	$\mathbf{y}^9_{(2)}$	$\mathbf{y}^9_{(3)}$	$\mathbf{y}^9_{(4)}$	$\mathbf{y}^9_{(5)}$	$\mathbf{y}^9_{(6)}$	$\mathbf{y}^9_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_5$	$\mathbf{y}_7$	$\mathbf{y}_3$	$\mathbf{y}_2$	$\mathbf{y}_1$	$\mathbf{y}_6$
	Mean	110.9	121.3	134.6	139.9	142.5	143.4	147.7
Y_{10}	Sort	$\mathbf{y}^{10}_{(1)}$	$\mathbf{y}^{10}_{(2)}$	$\mathbf{y}^{10}_{(3)}$	$\mathbf{y}^{10}_{(4)}$	$\mathbf{y}^{10}_{(5)}$	$\mathbf{y}^{10}_{(6)}$	$\mathbf{y}^{10}_{(7)}$
	Object	$\mathbf{y}_4$	$\mathbf{y}_3$	$\mathbf{y}_6$	$\mathbf{y}_5$	$\mathbf{y}_1$	$\mathbf{y}_7$	$\mathbf{y}_2$
	Mean	860.9	911.4	1056.2	1572.1	2009.1	2068.8	2168.0

Table B.4: Decrement values for the first stage.

Variable	(C_1^1, C_1^2)	$I(C_1^1)$	$I(C_1^2)$	$I(C_1) - I(C_1^1) - I(C_1^2)$
Y_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	1.16	6.85	4.52
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	1.91	2.85	**7.77**
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	4.03	1.12	7.38
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_7\})$	7.38	0.62	4.52
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_7\})$	9.86	0.00	2.67
Y_2	$(\{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_6\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_6\})$	1.66	8.30	2.57
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_6\})$	4.32	7.06	1.14
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_6\})$	6.39	4.11	2.02
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	7.47	0.31	4.75
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	10.24	0.00	2.28
Y_3	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\})$	0.00	10.70	1.83
	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\})$	0.13	8.31	4.08
	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\})$	1.12	4.03	7.38
	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5\}, \{\mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\})$	2.85	1.91	**7.77**
	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4\}, \{\mathbf{y}_6, \mathbf{y}_3\})$	7.47	0.31	4.75
	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_3\})$	10.49	0.00	2.03
	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\})$	0.00	10.70	1.83
Y_4	$(\{\mathbf{y}_4\}, \{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$	0.00	9.44	3.01
	$(\{\mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$	1.39	6.64	4.50
	$(\{\mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3\}, \{\mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$	1.91	2.85	**7.77**
	$(\{\mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_5\}, \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$	4.03	1.12	7.38
	$(\{\mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7\})$	7.62	0.92	3.98
	$(\{\mathbf{y}_4, \mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7\})$	9.86	0.00	2.67
Y_5	$(\{\mathbf{y}_4\}, \{\mathbf{y}_1, \mathbf{y}_6, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_3, \mathbf{y}_3\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_1\}, \{\mathbf{y}_6, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_3, \mathbf{y}_7\})$	3.81	8.15	0.56
	$(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_6\}, \{\mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_3, \mathbf{y}_7\})$	5.39	5.77	1.36
	$(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_6, \mathbf{y}_2\}, \{\mathbf{y}_5, \mathbf{y}_3, \mathbf{y}_7\})$	7.39	4.53	0.60
	$(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_6, \mathbf{y}_2, \mathbf{y}_5\}, \{\mathbf{y}_3, \mathbf{y}_7\})$	8.30	3.54	0.68
	$(\{\mathbf{y}_4, \mathbf{y}_1, \mathbf{y}_6, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_3\}, \{\mathbf{y}_7\})$	9.86	0.00	2.67
Y_6	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	1.16	6.85	4.52
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	1.91	2.85	**7.77**
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	4.03	1.12	7.38
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_7\})$	7.38	0.62	4.52
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_7\})$	9.86	0.00	2.67
Y_7	$(\{\mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4\})$	0.00	10.24	2.28
	$(\{\mathbf{y}_6, \mathbf{y}_3\}, \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4\})$	0.31	7.47	4.75
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4\})$	4.11	6.39	2.02
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_5, \mathbf{y}_4\})$	5.93	4.69	1.90
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5, \mathbf{y}_4\})$	8.30	1.66	2.57
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5\}, \{\mathbf{y}_4\})$	9.44	0.00	3.01
Y_8	$(\{\mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_2\})$	0.00	10.24	2.28
	$(\{\mathbf{y}_6, \mathbf{y}_3\}, \{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_2\})$	0.31	7.47	4.75
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_2\})$	1.91	2.85	**7.77**
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_2\})$	4.03	1.12	7.38
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_2\})$	8.31	0.13	4.08
	$(\{\mathbf{y}_6, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_1\}, \{\mathbf{y}_2\})$	10.80	0.00	1.73
Y_9	$(\{\mathbf{y}_4\}, \{\mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_3, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_6\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_3, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_6\})$	1.66	8.30	2.57
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_3, \mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_6\})$	4.69	5.93	1.90
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_3\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_6\})$	6.69	3.89	1.95
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_3, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_6\})$	8.75	2.88	0.90
	$(\{\mathbf{y}_4, \mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_3, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_6\})$	10.24	0.00	2.28
Y_{10}	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_7, \mathbf{y}_2\})$	0.00	9.44	3.09
	$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_7, \mathbf{y}_2\})$	1.16	6.85	4.52
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6\}, \{\mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_7, \mathbf{y}_2\})$	1.91	2.85	**7.77**
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5\}, \{\mathbf{y}_1, \mathbf{y}_7, \mathbf{y}_2\})$	4.03	1.12	7.38
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_1\}, \{\mathbf{y}_7, \mathbf{y}_2\})$	7.62	0.92	3.98
	$(\{\mathbf{y}_4, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_7\}, \{\mathbf{y}_2\})$	10.80	0.00	1.73

Table B.5: Decrement values at the second stage.

Variable	Cluster	$(C_u^1, C_u^2),\ u = 1, 2$	$I(C_u^1)$	$I(C_u^2)$	$I(C_u) - I(C_u^1) - I(C_u^2)$
Y_1	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	1.16	0.00	0.75
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_7\})$	1.08	0.62	1.15
		$(\{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_7\})$	1.69	0.00	1.16
Y_2	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	1.16	0.00	0.75
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_1\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_1\})$	1.08	0.62	1.15
		$(\{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_1\})$	2.42	0.00	0.43
Y_3	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_6, \mathbf{y}_3\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_3\})$	1.39	0.00	0.52
	C_2	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5\})$	0.00	2.42	0.43
		$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_5\})$	0.13	1.63	1.10
		$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5\})$	1.12	0.00	**1.73**
Y_4	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_6, \mathbf{y}_3\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_3\})$	1.39	0.00	0.52
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7\})$	1.32	0.92	0.61
		$(\{\mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7\})$	1.69	0.00	1.16
Y_5	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_6, \mathbf{y}_3\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_6\}, \{\mathbf{y}_3\})$	1.39	0.00	0.52
	C_2	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\})$	0.00	2.42	0.43
		$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_5, \mathbf{y}_7\})$	0.13	1.63	1.10
		$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5\}, \{\mathbf{y}_7\})$	1.69	0.00	1.16
Y_6	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	1.16	0.00	0.75
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_2, \mathbf{y}_1, \mathbf{y}_7\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_2\}, \{\mathbf{y}_1, \mathbf{y}_7\})$	1.08	0.62	1.15
		$(\{\mathbf{y}_5, \mathbf{y}_2, \mathbf{y}_1\}, \{\mathbf{y}_7\})$	1.69	0.00	1.16
Y_7	C_1	$(\{\mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	0.00	1.16	0.75
		$(\{\mathbf{y}_6, \mathbf{y}_3\}, \{\mathbf{y}_4\})$	0.31	0.00	1.60
	C_2	$(\{\mathbf{y}_1\}, \{\mathbf{y}_2, \mathbf{y}_7, \mathbf{y}_5\})$	0.00	2.42	0.43
		$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_5\})$	0.13	1.63	1.10
		$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5\})$	1.12	0.00	**1.73**
Y_8	C_1	$(\{\mathbf{y}_6\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	0.00	1.16	0.75
		$(\{\mathbf{y}_6, \mathbf{y}_3\}, \{\mathbf{y}_4\})$	0.31	0.00	1.60
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_1, \mathbf{y}_2\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_2\})$	1.63	0.13	1.10
		$(\{\mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_1\}, \{\mathbf{y}_2\})$	2.38	0.00	0.47
Y_9	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	1.16	0.00	0.75
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_2, \mathbf{y}_1\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_2, \mathbf{y}_1\})$	1.63	0.13	1.10
		$(\{\mathbf{y}_5, \mathbf{y}_7, \mathbf{y}_2\}, \{\mathbf{y}_1\})$	2.42	0.00	0.43
Y_{10}	C_1	$(\{\mathbf{y}_4\}, \{\mathbf{y}_3, \mathbf{y}_6\})$	0.00	0.31	1.60
		$(\{\mathbf{y}_4, \mathbf{y}_3\}, \{\mathbf{y}_6\})$	1.16	0.00	0.75
	C_2	$(\{\mathbf{y}_5\}, \{\mathbf{y}_1, \mathbf{y}_7, \mathbf{y}_2\})$	0.00	1.12	**1.73**
		$(\{\mathbf{y}_5, \mathbf{y}_1\}, \{\mathbf{y}_7, \mathbf{y}_2\})$	1.32	0.92	0.61
		$(\{\mathbf{y}_5, \mathbf{y}_1, \mathbf{y}_7\}, \{\mathbf{y}_2\})$	2.38	0.00	0.47

Table B.6: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_4\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_1$	$(\{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_4\})$	8.15	3.81	-2.52
$\mathbf{y}_{(2)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}, \{\mathbf{y}_2, \mathbf{y}_4\})$	8.25	3.74	-2.55
$\mathbf{y}_{(3)} \equiv \mathbf{y}_3$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_6, \mathbf{y}_7\}, \{\mathbf{y}_3, \mathbf{y}_4\})$	6.85	1.16	+1.16
$\mathbf{y}_{(4)} \equiv \mathbf{y}_5$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_6, \mathbf{y}_7\}, \{\mathbf{y}_5, \mathbf{y}_4\})$	8.30	1.66	-0.52
$\mathbf{y}_{(5)} \equiv \mathbf{y}_6$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_6, \mathbf{y}_4\})$	6.64	1.39	**+1.41**
$\mathbf{y}_{(6)} \equiv \mathbf{y}_7$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_6\}, \{\mathbf{y}_7, \mathbf{y}_4\})$	7.12	3.76	-1.44

Table B.7: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_4, \mathbf{y}_6\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_1$	$(\{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_4, \mathbf{y}_6\})$	5.77	5.39	-3.13
$\mathbf{y}_{(2)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_2, \mathbf{y}_4, \mathbf{y}_6\})$	5.86	5.30	-3.13
$\mathbf{y}_{(3)} \equiv \mathbf{y}_3$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$	2.85	2.91	**+3.27**
$\mathbf{y}_{(4)} \equiv \mathbf{y}_5$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_7\}, \{\mathbf{y}_5, \mathbf{y}_4, \mathbf{y}_6\})$	5.47	3.31	-0.75
$\mathbf{y}_{(5)} \equiv \mathbf{y}_7$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_4, \mathbf{y}_6\})$	4.94	5.96	-2.87

Table B.8: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_1$	$(\{\mathbf{y}_2, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$	2.42	6.27	-3.93
$\mathbf{y}_{(2)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_1, \mathbf{y}_5, \mathbf{y}_7\}, \{\mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$	2.38	6.09	-3.71
$\mathbf{y}_{(3)} \equiv \mathbf{y}_5$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_5, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$	1.16	4.03	-0.39
$\mathbf{y}_{(4)} \equiv \mathbf{y}_7$	$(\{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_5\}, \{\mathbf{y}_7, \mathbf{y}_3, \mathbf{y}_4, \mathbf{y}_6\})$	1.69	6.98	-3.91

Table B.9: $H_{(i)}$ values for $C_1^1 = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_7\}$ and $C_1^2 = \{\mathbf{y}_5\}$.

$\mathbf{y}_{(i)} \equiv \mathbf{y}_{i'}$	(TC^1, TC^2)	$I(TC^1)$	$I(TC^2)$	$H_{(i)}$
$\mathbf{y}_{(1)} \equiv \mathbf{y}_1$	$(\{\mathbf{y}_2, \mathbf{y}_7\}, \{\mathbf{y}_1, \mathbf{y}_5\})$	0.92	1.32	-1.13
$\mathbf{y}_{(2)} \equiv \mathbf{y}_2$	$(\{\mathbf{y}_1, \mathbf{y}_7\}, \{\mathbf{y}_2, \mathbf{y}_5\})$	0.62	1.08	-0.59
$\mathbf{y}_{(3)} \equiv \mathbf{y}_7$	$(\{\mathbf{y}_1, \mathbf{y}_2\}, \{\mathbf{y}_7, \mathbf{y}_5\})$	0.13	1.63	-0.64

Appendix C

Program Code in R software

```
        ################################################################
        #                                                              #
        #       Cluster Analysis for Histogram-valued Data             #
        #                     By Jaejik Kim                            #
        #                   R : version 2.5.0                          #
        #                                                              #
        ################################################################

#==== Create a Histogram-valued Dataset from Classical Dataset.=====

histdata=function(x,v,c)
  # x : data frame
  # v : columns to be histogram-valued data
  # c : column for a category variable
{
  a=levels(factor(x[,c]))
  m=length(a)
  p=length(v)
  nm=colnames(x)[v]

  # Function to make ( , , , ) ------------------
  addc=function(q)
    # q : vector
  {
    k=length(q)
    f=q[1]
    if (k>=2)
    {
      for (i in 1:(k-1))
      {
        f=paste(f,",",q[i+1])
      }
    }
    f=paste("(",f,")")
    return(f)
  }
  #-----------------------------------------------

  br=matrix(nrow=m,ncol=p)
  rf=matrix(nrow=m,ncol=p)
  for (i in 1:m)
  {
    y=x[x[,c]==a[i],]
    for (j in 1:p)
    {
      z=hist(y[,v[j]],plot=FALSE)
      b=z$breaks
      r=z$counts/sum(z$counts)
      br[i,j]=addc(b)
      rf[i,j]=addc(r)
    }
  }
  colnames(br)=nm
  rownames(br)=a
  colnames(rf)=nm
  rownames(rf)=a
  result=list(br,rf)
  names(result)=c("br","rf")
  return(result)
}

#=================== Data Transformation ====================

hist.data = function(dt,fq,ls="M")
  # dt : Break points data for intervals
  # fq : Relative frequency for intervals
  # ls : the type of the length of transformed subintervals
  #     (M :minimum length, A : Average length)
{
  dt=as.matrix(dt)
  fq=as.matrix(fq)
  nc1=nchar(dt)
  nc2=nchar(fq)
  dt=substr(dt,2,nc1-1)
  fq=substr(fq,2,nc2-1)
  rn=rownames(dt)           # rn : name of obs.
  cn=colnames(dt)           # cn : name of variables

  m=length(dt[,1])          # m  : number of observations
```

```
p=length(dt[1,])          # p  : number of variables
q=vector(length=p)        # q  : maximum number of intervals
SINTV=vector(length=p)    # SINTV : minimum or average length
                          #         of intervals
for (i in 1:p)            # Detect the maximum number of intervals
{                         # for each variable
    g=0                   # g : initial value for maximum number
                          #     of intervals
    INTV=vector(length=m)
  for (j in 1:m)
  {  z1=unlist(strsplit(dt[j,i], ",", fixed = TRUE))
     z1=as.numeric(z1)
     n1=length(z1)
     if (n1 > g) g=n1
     INTV[j]=abs(z1[2]-z1[1])
  }
  q[i]=g
  if (ls=="M") SINTV[i]=min(INTV)
  if (ls=="A") SINTV[i]=mean(INTV)
}

cq=cumsum(q)              # Assign the data to matrix
dt1=matrix(nrow=m,ncol=sum(q))
fq1=matrix(nrow=m,ncol=sum(q)-p)
for (i in 1:p)
{ for (j in 1:m)
  {  z1=unlist(strsplit(dt[j,i], ",", fixed = TRUE))
     z1=as.numeric(z1)
     n1=length(z1)
     z2=unlist(strsplit(fq[j,i], ",", fixed = TRUE))
     z2=as.numeric(z2)
     n2=length(z2)
     n3=n1-1
     if (n3!=n2)         # Error massage : check the data out!
      stop ("The number of break points and relative
             frequencies do NOT match")
     if (i==1)
     {dt1[j,1:n1]=z1
      fq1[j,1:n2]=z2}
     if (i>1)
     {cq1=cq[i-1]+1
      cq2=cq1+n1-1
      dt1[j,cq1:cq2]=z1
      cq3=cq[i-1]-i+2
      cq4=cq3+n2-1
      fq1[j,cq3:cq4]=z2}
  }
}

min1=vector(length=p)     # min1 : minimum value of
                          #        histogram-valued variable
max1=vector(length=p)     # max1 : maximum value of
                          #        histogram-valued variable
for (i in 1:p)
{  if (i==1)
   {min1[i]=min(dt1[,1:cq[i]],na.rm=T)
    max1[i]=max(dt1[,1:cq[i]],na.rm=T) }
   if (i>1)
   {cq1=cq[i-1]+1
    min1[i]=min(dt1[,cq1:cq[i]],na.rm=T)
    max1[i]=max(dt1[,cq1:cq[i]],na.rm=T) }
}

nq=round(abs(max1-min1)/SINTV)+1
                # Number of new intervals for each variable
cnq=cumsum(nq)
new=0
for (i in 1:p)                    # Construct new intervals
{ new1=seq(min1[i],max1[i],length=nq[i])
  new=append(new,new1)
}
new=new[2:length(new)]
cfq=cumsum(q-1)

for (j in 1:m)
{ new.fq1=0
  for (i in 1:p)
  {  if (i==1)
     {z1=dt1[j,1:cq[i]]
      z2=fq1[j,1:cfq[i]]
      z3=new[1:cnq[i]] }
     if (i>1)
     {cq1=cq[i-1]+1
      cq2=cfq[i-1]+1
      cq3=cnq[i-1]+1
      z1=dt1[j,cq1:cq[i]]         # z1 : Old break points
      z2=fq1[j,cq2:cfq[i]]        # z2 : Old relative frequencies
      z3=new[cq3:cnq[i]] }        # z3 : New break points
    n1=length(z3)-1
    n2=length(na.omit(z1))
```

```
z4=vector(length=n1)          # z4 : New relative frequencies
n=1
for (k in 1:n1)
{if (n==1)
  {if (z1[n]>=z3[k+1]) z4[k]=0
   if (z1[n]>=z3[k] & z1[n]<z3[k+1] & z1[n+1]>=z3[k+1])
       {z4[k]=(abs(z3[k+1]-z1[n])/abs(z1[n+1]-z1[n]))*z2[n]}
   if (z1[n]>=z3[k] & z1[n+1]<z3[k+1])
      {
       if (z1[n2]>z3[k+1])
       {t=n
        while (z1[t+1]<z3[k+1] & t<=n2)
        {t=t+1}
        z4[k]=sum(z2[n:(t-1)])+(abs(z3[k+1]-z1[t])
              /abs(z1[t+1]-z1[t]))*z2[t]}
       if (z1[n2]<=z3[k+1]) z4[k]=sum(z2[n:(n2-1)])
      }
   if (z1[n]<z3[k] & z1[n+1]>=z3[k+1])
      {z4[k]=(abs(z3[k+1]-z3[k])/abs(z1[n+1]-z1[n]))*z2[n]}

 if (n>1 & n<n2)
  {
   if (z1[n]>=z3[k] & z1[n]<z3[k+1] & z1[n+1]>=z3[k+1])
   {z4[k]=((abs(z1[n]-z3[k])/abs(z1[n]-z1[n-1]))*z2[n-1])
    +((abs(z3[k+1]-z1[n])/abs(z1[n+1]-z1[n]))*z2[n])}
   if (z1[n]>=z3[k] & z1[n+1]<z3[k+1])
      {
       if (z1[n2]>z3[k+1])
       {t=n
        while (z1[t+1]<z3[k+1])
        {t=t+1}
        z4[k]=((abs(z1[n]-z3[k])/abs(z1[n]-z1[n-1]))*z2[n-1])
               +sum(z2[n:(t-1)])+(abs(z3[k+1]-z1[t])
               /abs(z1[t+1]-z1[t]))*z2[t]}
       if (z1[n2]<=z3[k+1])
        {z4[k]=((abs(z1[n]-z3[k])/abs(z1[n]-z1[n-1]))*z2[n-1])
                 +sum(z2[n:(n2-1)])}
       }
   if (z1[n]<z3[k] & z1[n+1]>=z3[k+1])
      {z4[k]=(abs(z3[k+1]-z3[k])/abs(z1[n+1]-z1[n]))*z2[n] }
   }

 if (n==n2)
 {if (z1[n]>z3[k]) z4[k]=(abs(z1[n]-z3[k])/abs(z1[n]-z1[n-1]))*z2[n-1]
  if (z1[n]<=z3[k]) z4[k]=0  }
```

```
       if (n>n2 | is.na(z1[n])) z4[k]=0

        # Count for real interval values(z1)
        if (n==n2) n=n+1
        if (n<n2 & k<n1)
         {if (z1[n]>=z3[k]&z1[n]<z3[k+1]&z1[n+1]>=z3[k+1]&z1[n+1]<z3[k+2])
             {n=n+1}}
        if (n<n2)
         {if (z1[n]>=z3[k] & z1[n+1]<z3[k+1] & z1[n2]>z3[k+1]) n=t+1
          if (z1[n]>=z3[k] & z1[n+1]<z3[k+1] & z1[n2]<=z3[k+1]) n=n2}
        if (n<n2 & k<n1)
         {if (z1[n]<z3[k] & z1[n+1]>=z3[k+1] & z1[n+1]<z3[k+2]) n=n+1}
      }         # End of for (k)
      new.fq1=append(new.fq1,z4)
    }           # End of for (i)
    if (j==1) new.fq=new.fq1
    if (j>1) new.fq=rbind(new.fq,new.fq1)
  }             # End of for (j)
  n=ncol(new.fq)
  new.fq=new.fq[,2:n]
  rownames(new.fq)=rn
  names(cnq)=cn

  result=list(new,new.fq,cnq)
  names(result)=c("bp","rfq","cnq")
  return(result)
}

#===================== Dissimilarity Measure =====================

dissim <- function
          (hdata,measure="E",gamma=0.5,Normal=TRUE,weight=FALSE,wj)
  # hdata   : object of fuction "hist.data"
  # measure : Dissimilarity measures
  #           (Gowda-Diday(GD),city block(CB),Euclidean(E),CDF)
  # gamma   : Coefficient for Ichino-Yaguchi measure.(Default=0.5)
  # Normal  : Whether it gives a normalized value for Ichino-Yaguchi
  #           or CDF measure or not(Default=TRUE)
  # weight  : Whether it gives weights
  # wj      : weight for I-Y measure
  #           (Default: weight=TRUE -> wj=1/p, weight=FALSE -> wj=1)
{
    bp=hdata$bp
    rfq=hdata$rfq
    cnq=hdata$cnq
```

```
m=nrow(rfq)
p=length(cnq)
cq=cnq-1:p
d.temp=array(0,dim=c(m,m,p)) # Distance matrix for each variable
MU=array(0,dim=c(m,m,p))     # Mean of Union
rn=rownames(rfq)
                 # (To obtain cut points in Monothetic algorithm)
cn=names(cnq)

mean.hv <- function (bp,f)   # Mean of Histogram for an obs.
{   n=length(bp)
    a=bp[1:(n-1)]
    b=bp[2:n]
    m1=sum(((b+a)*f)/2)
    return(m1)  }

sd.hv <- function (bp,f)    # Standard deviation for an histogram obs.
{   n=length(bp)
    a=bp[1:(n-1)]
    b=bp[2:n]
    sf=sum(f)
    if (sf!=0) f1=f/sum(f)
    if (sf==0) f1=f
    M=sum(((b+a)*f1)/2)
    v=sqrt(sum((((a-M)^2+(a-M)*(b-M)+(b-M)^2)/3)*f))
    return(v)   }


for (i in 1:m)
  {for (j in 1:m)
    {if (j>i)
        {
     for (k in 1:p)
       {if (k==1)
        {f0=rfq[,1:cq[1]]
         f1=rfq[i,1:cq[1]]
         f2=rfq[j,1:cq[1]]
         z=bp[1:cnq[1]] }
        if (k>1)
        {cq1=cq[k-1]+1
         cq2=cnq[k-1]+1
         f0=rfq[,cq1:cq[k]]
         f1=rfq[i,cq1:cq[k]]
         f2=rfq[j,cq1:cq[k]]
         z=bp[cq2:cnq[k]]   }
```

```
 t=length(f1)               # The number of subintervals
 Psi=z[t+1]-z[1]            # The length spanned by obs

 fx=cbind(f1,f2)
 fu=apply(fx,1,max)
 fi=apply(fx,1,min)

 # Mean of Union for cut points______________________
 fu.cut=fu/sum(fu)
 MU[i,j,k]=mean.hv(z,fu.cut)
 MU[j,i,k]=MU[i,j,k]
 # _______________________________________________End

 if (measure!="CDF")
 {
   s1=sd.hv(z,f1)               # Standard deviation for A
   s2=sd.hv(z,f2)               # Standard deviation for B
 su=sd.hv(z,fu)
 si=sd.hv(z,fi)
 }

 # Calculation for Measures(GD, CB, E, CDF)

 if (measure=="GD")    # Gowda-Diday Dissimilarity Measure
{
  d.temp[i,j,k]=abs(s1-s2)/(s1+s2)+(s1+s2-2*si)/(s1+s2)
                +abs(mean.hv(z,f1)-mean.hv(z,f2))/Psi
  d.temp[j,i,k]=d.temp[i,j,k]
    if (k==p) D=apply(d.temp,c(1,2),sum)
}

 if (measure=="CB" | measure=="E" | missing(measure))
{
 phi=su-si+gamma*(2*si-s1-s2)
                      # Ichino-Yaguchi(r=0.5, 0=<r=<0.5)

 if (Normal==T)       # Normalized Ichino-Yaguchi measure
 {
   a1=z[1]
   b1=z[2]
   at=z[t]
   bt=z[t+1]
   V=sqrt((5*(a1^2+b1^2+at^2+bt^2)+2*(a1*b1+at*bt)
           -6*(a1*at+a1*bt+at*b1+b1*bt))/24)
           phi=phi/V
         }
         if (weight==F) wj=rep(1,p)
         if (weight==T & missing(wj)) wj=rep(1/p,p)

         if (measure=="CB")       # City Block Distance
           {
             d.temp[i,j,k]=wj[k]*phi
             d.temp[j,i,k]=d.temp[i,j,k]
             if (k==p) D=apply(d.temp,c(1,2),sum)
           }

         if (measure=="E" | missing(measure))
           {                # Normalized Euclidean Distance
             d.temp[i,j,k]=wj[k]*phi^2
             d.temp[j,i,k]=d.temp[i,j,k]
                 if (k==p) D=sqrt(apply(d.temp,c(1,2),sum))
           }
        }

         if (measure=="CDF")       # CDF Distance
        {b=length(z)
         z1=z[1:(b-1)]
         z2=z[2:b]
         fa=cumsum(f1)
         fb=cumsum(f2)
         d.temp[i,j,k]=sum(abs(z2-z1)*abs(fa-fb))
         d.temp[j,i,k]=d.temp[i,j,k]
         if (Normal==T)
           {
             d.temp[i,j,k]=d.temp[i,j,k]/Psi
             d.temp[j,i,k]=d.temp[j,i,k]/Psi
           }
                if (k==p) D=apply(d.temp,c(1,2),sum)
        }
       }
      }
    }
  }
dp=apply(d.temp,3,sum)
dp=dp/sum(dp)
rownames(D)=rn
colnames(D)=rn
names(dp)=cn
dp=sort(dp,decreasing=TRUE)
```

```
    result=list(D,d.temp,MU,dp)
    names(result)=c("dsm","dmv","MU","impt")
               # dsm  : distance matrix
               # dmv  : distance matrix for each variable
               # MU   : mean matrix of union (cut points)
               # impt : relatively important degree for each variable
    return(result)
}

#============= Divisive Clustering Method (Monothetic)===========
 mdivclust <- function(hdata, dsmty)
  # hdata   : object of fuction "hist.data"
  # dsmty   : object of fuction "dissim"
{
    bp=hdata$bp
    rfq=hdata$rfq
    cnq=hdata$cnq
    dsm=dsmty$dsm
    dmv=dsmty$dmv
    MU=dsmty$MU

    m=nrow(dsm)                    # # of obs.
    rn=rownames(dsm)
    cn=names(cnq)
    p=length(cnq)
    cq=cnq-1:p
    mn=1:m

    # Mean function for Histogram-valued data____________________
    mean.hv <- function (bp,f)     # Mean of Histogram for an obs.
    {   n=length(bp)
        a=bp[1:(n-1)]
        b=bp[2:n]
        m1=sum(((b+a)*f)/2)
        return(m1)  }
    #____________________ End of Mean Function

    ### Mean for each obs and variable ###
    hmean=matrix(nrow=m,ncol=p)
    for (j in 1:p)
    {  for (i in 1:m)
        {if (j==1)
           {f=rfq[i,1:cq[1]]
            z=bp[1:cnq[1]]}
         if (j>1)
        {cq1=cq[j-1]+1
         cq2=cnq[j-1]+1
         f=rfq[i,cq1:cq[j]]
         z=bp[cq2:cnq[j]]   }
     hmean[i,j]=mean.hv(z,f)
    }
}
rownames(hmean)=mn
colnames(hmean)=cn

### Divisive clustering(Monothetic) ###

#___________Function for the part of the Within-Cluster variance
# dsm : Dissimilarity matrix, sn : elements of cluster
WCV <- function (dsm,sn)
    {  mk=length(sn)
       m=nrow(dsm)
       if (mk==1) IC=0
       if (mk>1)
       {    IC=0
        for (i in 1:mk)
        {  for (j in 1:mk)
           {  if (j>i)
            {  t1=sn[i]
               t2=sn[j]
               IC=IC+(dsm[t1,t2])^2 }   } }
           IC=IC/(m*mk)   }
       return(IC)   }
  # End of the function WCV __________

m1=2*(m-1)+1
clus=matrix(nrow=m1,ncol=m) # Record for grouping
clus[1,]=mn                 # 1st row = all the observations
div=matrix(0,nrow=(m-1),ncol=6)
  # column 1 : origin, column 2-3 : divided cluster
  # column 4 : Total within-cluster variation
  # column 5 : Total between-cluster variation
  # column 6 : total variation explained by
  #            the differences between clusters

group=1                  # Rows of matrix 'clus'
question=vector(length=(m-1))
WE=WCV(dsm,mn)           # Total within-cluster variance for set E
for (i in 1:(m-1))       # i : Total number of steps
{  r=length(group)
```

```
WCk0=0
for (j in 1:p)        # j : Total number of variables
{    hm=hmean[,j]
 for (k in 1:r)       # k : Total number of clusters at each step
 {  a1=group[k]
    a2=na.omit(clus[a1,])
    a3=length(a2)
    if (a3 > 1)
    {  hmp=0             # Sorting for the mean
       for (b in 1:a3)
        {  hmp=rbind(hmp,hm[names(hm)==a2[b]])  }
       hmp=hmp[2:(a3+1)]
       names(hmp)=a2
       sn=as.numeric(names(sort(hmp)))
     IC=WCV(dsm,sn)

       for (l in 1:(a3-1))
     {  sn1=sn[1:l]
        sn2=sn[(l+1):a3]
        IC1=WCV(dsm,sn1)
        IC2=WCV(dsm,sn2)
        WCk=IC-IC1-IC2

        if (round(WCk0,8) == round(WCk,8))
        {
         dv1=dmv[snl1,snl2,vr]
         dv2=dmv[sn[l],sn[l+1],j]

         if (dv2 > dv1)
         {
            clus1=sn1
            clus2=sn2
            div1=a1
            div2=2*i
            div3=2*i+1
            vr=j
            snl1=sn[l]
            snl2=sn[l+1]
            dv1=dv2
            grp=append(group,c(2*i,2*i+1),after=k)
            locat=which(grp==a1)
            grp=grp[-(grp=locat)]
            variable=cn[j]
            cut=MU[snl1,snl2,vr]
            cut=round(cut,4)
            }
           }

           if (round(WCk,8) > round(WCk0,8))
           {  clus1=sn1
            clus2=sn2
            div1=a1
            div2=2*i
            div3=2*i+1
            WCk0=WCk
            vr=j
            snl1=sn[l]
            snl2=sn[l+1]
            grp=append(group,c(2*i,2*i+1),after=k)
            locat=which(grp==a1)
            grp=grp[-(grp=locat)]
            variable=cn[j]
            cut=MU[snl1,snl2,vr]
            cut=round(cut,4)
           }
        }
       }
    }
   }
l1=length(clus1)
l2=length(clus2)
clus[(2*i),1:l1]=clus1
clus[(2*i+1),1:l2]=clus2
div[i,1]=div1
div[i,2]=div2
div[i,3]=div3
question[i]=paste(variable,"<=",round(cut,2))
group=grp

WP=0
r=length(group)
for (o in 1:r)
{
    a1=group[o]
    a2=na.omit(clus[a1,])
    WP=WP+WCV(dsm,a2)
}
div[i,4]=WP
div[i,5]=WE-WP
div[i,6]=(WE-WP)/WE
```

```
    }
    colnames(div)=c("Origin","YES","NO","Within","Between","Explained")
    rownames(div)=paste("step",1:(m-1))

    names(rn)=mn
    result=list(div,clus,rn,question)
    names(result)=c("div","cluster","obs","question")
    return(result)
}

#========= Divisive Clustering Method (Polyhetic)============

pdivclust <- function(dsmty)
  # dsmty   : dissimilarity matrix
{
    dsm=dsmty$dsm
    m=nrow(dsm)             # # of obs.
    rn=rownames(dsm)
    mn=1:m

    ### Divisive clustering(Polythetic) ###

    #___________Function for the part of the Within-Cluster variance
    # dsm : Dissimilarity matrix, sn : vector for elements of cluster
    WCV <- function (dsm,sn)
        {  mk=length(sn)
           m=nrow(dsm)
           if (mk==1) IC=0
           if (mk>1)
           {    IC=0
            for (i in 1:mk)
            {  for (j in 1:mk)
               {  if (j>i)
                {  t1=sn[i]
                   t2=sn[j]
                   IC=IC+(dsm[t1,t2])^2 }   } }
               IC=IC/(m*mk)   }
           return(IC)   }
      # End of the function WCV __________

    #___________Function getting average dissimilarity or distance
    # dsm : Dissimilarity matrix, sn : vector for elements of cluster
    MAVD <- function (dsm,sn)
    {
      n=nrow(dsm)
  mk=length(sn)
  if (mk==1)
  {
    mavd=0
    seed=1
  }
  if (mk>1)
  {
    dsm1=dsm[sn,sn]
    avd=apply(dsm1,1,sum)/((mk-1)*n)
    mavd=max(avd)
    seed=which(avd==mavd)
    seed=sn[seed[1]]
  }
  result=list(mavd,seed)
  names(result)=c("maxd","seed")  # maxd : maximum average distance
  return(result)                  # seed : object that has maxd
}
  # End of the function MAVD______

m1=2*(m-1)+1
clus=matrix(nrow=m1,ncol=m) # Record for grouping
clus[1,]=mn                 # 1st row = all the observations
div=matrix(0,nrow=(m-1),ncol=6)
  # column 1 : origin, column 2-3 : divided cluster
  # column 4 : Total within-cluster variation
  # column 5 : Total between-cluster variation
  # column 6 : total variation explained
  #            by the differences between
  clusters
group=1                  # Rows of matrix ’clus’
WE=WCV(dsm,mn)           # Total within-cluster variance for set E
for (i in 1:(m-1))       # i : Total number of steps
{
  r=length(group)
  maxg=1:r
  lmax=1:r
  for (k in 1:r)
  {
    a1=group[k]
    sn=na.omit(clus[a1,])
    ma=MAVD(dsm,sn)
    maxg[k]=ma$maxd
    lmax[k]=ma$seed
  }
```

```
id=which(maxg==max(maxg))
id=id[1]
mg=group[id] # mg : cluster which has maximum average distance
a1=na.omit(clus[mg,])
a2=length(a1)
if (a2>1)
{
  C2=lmax[id]             # C2 : Splinter group
  a3=which(a1==C2)
  C1=a1[-a3]              # C1 : Main group
}

if (a2>2)
{
  MIC12=1
  while(MIC12>0)
  {
  b=length(C1)
  DIC=vector(length=b)
  IC=WCV(dsm,C1)+WCV(dsm,C2)
  for (j in 1:b)
  {
    TC2=c(C1[j],C2)
    TC1=C1[-j]
    DIC[j]=IC-WCV(dsm,TC1)-WCV(dsm,TC2)
  }
  MIC12=max(DIC)
  if (MIC12>0)
  {
    a4=which(DIC==MIC12)
    C2=c(C1[a4],C2)
    C1=C1[-a4]
  }
  }
}
l1=length(C1)
l2=length(C2)
clus[(2*i),1:l1]=C1
clus[(2*i+1),1:l2]=C2
div[i,1]=mg
div[i,2]=2*i
div[i,3]=2*i+1

group=append(group,c(2*i,2*i+1),after=id)
locat=which(group==mg)
```

```
      group=group[-(group=locat)]

      WP=0
      r=length(group)
      for (o in 1:r)
      {
        a1=group[o]
        a2=na.omit(clus[a1,])
        WP=WP+WCV(dsm,a2)
      }
      div[i,4]=WP
      div[i,5]=WE-WP
      div[i,6]=(WE-WP)/WE
    }

    colnames(div)=c("Origin","Main","Splinter","Within",
                    "Between","Explained")
    rownames(div)=paste("step",1:(m-1))

    names(rn)=mn
    result=list(div,clus,rn)
    names(result)=c("div","cluster","obs")
    return(result)
}


#=================================================================

# Calculate a Validity Index(DB*, DI)

CVI <- function(clus, dsmty)
  # dsmty (object) : dissimilarity measure
  # clus  (object) : Monothetic Method
  # pclus (object) : Polythetic Method

{
    dsm=dsmty$dsm
    clust=clus$cluster
    merge=clus$div[,1]
    yn=clus$div[,2:3]
    obs=clus$obs
    m=length(obs)

    #___________Function for the part of the Within-Cluster variance
    # dsm : Dissimilarity matrix, sn : vector for elements of cluster
```

```
WCV <- function (dsm,sn)
    {  mk=length(sn)
       m=nrow(dsm)
       if (mk==1) IC=0
       if (mk>1)
       {    IC=0
        for (i in 1:mk)
        {  for (j in 1:mk)
           {  if (j>i)
            {  t1=sn[i]
               t2=sn[j]
               IC=IC+(dsm[t1,t2])^2 }   } }
           IC=IC/(m*mk)   }
       return(IC)   }
  # End of the function WCV __________

# Calculate Cluster Validity Indeces (DB, DB*, DI, CS, Vsv)

result=matrix(nrow=2,ncol=(m-2))
rownames(result)=c("DB", "DI")
colnames(result)=2:(m-1)

for (k in 1:(m-2))
{

  # Elements of each cluster at each stage

  a=yn[k,]
  temp=clust[a,]
  rownames(temp)=a
  if (k==1) Pr=temp
  if (k>=2)
  {
    Pr=rbind(Pr,temp)
    a1=which(rownames(Pr)==merge[k])
    Pr=Pr[-a1,]
  }

  r=nrow(Pr)
  IC=vector(length=r)
  for (i in 1:r)
  {
    temp0=na.omit(Pr[i,])
    IC[i]=WCV(dsm,temp0)
  }
  t=1
  ID=matrix(nrow=sum((r-1):1),ncol=2)
  ICU=vector(length=sum((r-1):1))
  T.DB=vector(length=sum((r-1):1))
  T.DB1=vector(length=sum((r-1):1))
  for(i in 1:r)
  {
    for(j in 1:r)
    {
    if (i<j)
    {
      ID[t,]=c(i,j)
      Ci=na.omit(Pr[i,])
      Cj=na.omit(Pr[j,])
      iUj=c(Ci,Cj)
      ICU[t]=WCV(dsm,iUj)-IC[i]-IC[j]
      T.DB1[t]=IC[i]+IC[j]
      t=t+1
    }
    }
  }

  ID1=ID[,1]
  ID2=ID[,2]

  DB1=vector(length=r)
  DI=vector(length=r)
  mxIC=max(IC)

  for (i in 1:r)
  {
    t1=which(ID1==i)
    t2=which(ID2==i)
    t=c(t1,t2)

    DB1[i]=(max(T.DB1[t]))/(min(ICU[t]))
    DI[i]=min(ICU[t]/mxIC)
  }

  result[1,k]=(sum(DB1))/r
  result[2,k]=min(DI)
}

return(result)
```

```
}

#========================== Dandrogram==============================

divplot <- function(clus,method,tsize=1,tit)
    # clus  : object for function "divclust"
    # mehtod    : "M" monothetic , "P" : polythetic
    # tsize : text size in the plot.
    # tit       : title of the plot
{
    clust=clus$cluster
    merge=clus$div[,1]
    yn=clus$div[,2:3]
    obs=clus$obs
    if (method=="M") qs=clus$question
    m=length(obs)
    m1=m-1

    ### Get the order for the observations in the dandrogram ###
    for (i in 1:m1)
    {
       if (i==1)
       {    a1=na.omit(clust[(2*i),])
        a2=na.omit(clust[(2*i+1),])
        odr=append(a1,a2)   }
       if (i>1)
       {    a1=na.omit(clust[(2*i),])
        a2=na.omit(clust[(2*i+1),])
        temp=append(a1,a2)
        m2=length(temp)
        loc=0
          for (j in 1:m2)
        {  a3=which(odr==temp[j])
           loc=rbind(loc,a3)    }
        loc=loc[2:(m2+1)]
        b1=min(loc)
        b2=max(loc)
        odr[b1:b2]=temp    }
    }

    ########## Drow the plot ##########

    x=c(0,0)
    y=c(0,0)
    plot(x,y,xlim=c(-10,115),ylim=c(-15,110),type="l",
     axes=FALSE,xlab="",ylab="")
xy=seq(0,100,length=m)
vert=m:2
node=matrix(0,nrow=m1,ncol=2)

for (i in m1:1)
{  a1=na.omit(clust[yn[i,1],])
   a2=na.omit(clust[yn[i,2],])
   if (length(a1)==1 & length(a2)==1)
   {
    b1=which(odr==a1)
    b2=which(odr==a2)
    x1=xy[b1]
    x2=xy[b2]
    h1=xy[vert[i]]
    lines(c(x1,x1),c(0,h1))
    lines(c(x2,x2),c(0,h1))
    lines(c(x1,x2),c(h1,h1))
   }
   if (length(a1)==1 & length(a2)>1)
   {
    b1=which(odr==a1)
    b2=which(merge==yn[i,2])
    x1=xy[b1]
    x2=node[b2,1]
    h1=xy[vert[i]]
    h2=node[b2,2]
    lines(c(x1,x1),c(0,h1))
    lines(c(x2,x2),c(h2,h1))
    lines(c(x1,x2),c(h1,h1))
   }
   if (length(a1)>1 & length(a2)==1)
   {
    b1=which(merge==yn[i,1])
    b2=which(odr==a2)
    x1=node[b1,1]
    x2=xy[b2]
    h1=xy[vert[i]]
    h2=node[b1,2]
    lines(c(x1,x1),c(h2,h1))
    lines(c(x2,x2),c(0,h1))
    lines(c(x1,x2),c(h1,h1))
   }
   if (length(a1)>1 & length(a2)>1)
   {
        b1=which(merge==yn[i,1])
        b2=which(merge==yn[i,2])
        x1=node[b1,1]
        x2=node[b2,1]
        h1=xy[vert[i]]
        h2=node[b1,2]
        h3=node[b2,2]
        lines(c(x1,x1),c(h2,h1))
        lines(c(x2,x2),c(h3,h1))
        lines(c(x1,x2),c(h1,h1))
       }
       node[i,1]=(x1+x2)/2
       node[i,2]=h1
    }
    lines(c(node[1,1],node[1,1]),c(node[1,2]+5,node[1,2]))
    odrname=vector(length=m)

    for (i in 1:m)
    {  odrname[i]=obs[odr[i]]   }

    y=rep(-3,m)
    text(x=xy,y=y,labels=odrname,cex=tsize)
    node1=node[,1]
    node2=node[,2]+3
    if (!missing(tit)) title(main=tit)
    if (method=="M")
    {
      text(x=node1,y=node2,labels=qs,cex=tsize,col=2)
      legend(x=30,y=-10,legend="Left : YES , Right : NO",cex=0.7)
    }
    if (method=="P") legend(x=30,y=-10,
                     legend="Right : Splinter group",cex=0.7)
}
```