

# ScholarWorks@GSU

## Classification For 2011-2012 Bangladesh Integrated Household Survey By Iterative Clustering Technique

|               |   |
|---------------|---|
| Authors       | Zhou, Wen   |
| Citation      | Zhou, Wen. "Classification For 2011-2012 Bangladesh Integrated Household Survey By Iterative Clustering Technique." 2013. Thesis, Georgia State University. <a href="https://doi.org/10.57709/4311238">https://doi.org/10.57709/4311238</a> |
| DOI           | <a href="https://doi.org/10.57709/4311238">https://doi.org/10.57709/4311238</a>   |
| Download date | 2026-06-14 03:30:58   |
| Link to Item  | <a href="https://hdl.handle.net/20.500.14694/10431">https://hdl.handle.net/20.500.14694/10431</a>   |

CLASSIFICATION FOR 2011-2012 BANGLADESH INTEGRATED HOUSEHOLD SURVEY BY ITERATIVE  
CLUSTERING TECHNIQUE

by

WEN ZHOU

Under the Direction of Xin Qi

ABSTRACT

In this project, the raw data from a survey, 2011-2012 Bangladesh Integrated Household Survey, is cleaned. Based on the research purpose of the collaborator, important variables are extracted and principal component analysis is used to form a new data set. The iterative clustering technique is applied to the new data set to classify the households involved in the survey into different categories. The categories are interpreted as reflecting the different economic activities in Bangladesh.

INDEX WORDS: Data clustering, Principal component analysis, Agriculture, Livestock, Rural development, Fisheries, Integrated farming, Bangladesh

CLASSIFICATION FOR 2011-2012 BANGLADESH INTEGRATED HOUSEHOLD SURVEY BY ITERATIVE  
CLUSTERING TECHNIQUE

by

WEN ZHOU

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2013

Copyright by  
Wen Zhou  
2013

CLASSIFICATION FOR 2011-2012 BANGLADESH INTEGRATED HOUSEHOLD SURVEY BY ITERATIVE CLUSTERING TECHNIQUE

by

WEN ZHOU

Committee Chair: Xin Qi

Committee: Ruiyan Luo

Yi Jiang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2013

## **ACKNOWLEDGEMENTS**

I take this opportunity to express my profound gratitude and deep regards to my guide (Professor Xin Qi) for his exemplary guidance, monitoring and constant encouragement through the course of this thesis. The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark.

Lastly, I thank almighty, my parents, husband, and my son for their constant encouragement without which this assignment would not be possible.

## TABLE OF CONTENTS

|   |                                      |
|---|--------------------------------------|
| <b>ACKNOWLEDGEMENTS .....</b>                         | <b>iError! Bookmark not defined.</b> |
| <b>LIST OF TABLES .....</b>                           | <b>vi</b>                            |
| <b>LIST OF FIGURES.....</b>                           | <b>vii</b>                           |
| <b>1 INTRODUCTION .....</b>                           | <b>1</b>                             |
| <b>1.1 Purpose of the Study .....</b>                 | <b>2</b>                             |
| <b><i>1.1.1 Background of the survey .....</i></b>    | <b>2</b>                             |
| <b><i>1.1.2 Purpose of the study.....</i></b>         | <b>2</b>                             |
| <b>2 DATA CLEANING .....</b>                          | <b>4</b>                             |
| <b>3 CLASSIFICATION OF HOUSEHOLDS.....</b>            | <b>5</b>                             |
| <b>3.1 Variables grouping and PCA.....</b>            | <b>5</b>                             |
| <b><i>3.1.1 Variables grouping .....</i></b>          | <b>5</b>                             |
| <b><i>3.1.2 Principal component analysis.....</i></b> | <b>7</b>                             |
| <b><i>3.1.3 New data set.....</i></b>                 | <b>10</b>                            |
| <b>3.2 Iterative clustering .....</b>                 | <b>10</b>                            |
| <b>3.3 Result .....</b>                               | <b>16</b>                            |
| <b>4 CONCLUSIONS .....</b>                            | <b>21</b>                            |
| <b>REFERENCES .....</b>                               | <b>22</b>                            |
| <b>APPENDIX: R SCRIPT.....</b>                        | <b>23</b>                            |

**LIST OF TABLES**

|   |    |
|---|----|
| Table 1 Clustering criterion.....                         | 11 |
| Table 2 1 <sup>st</sup> time observations allocation..... | 13 |
| Table 3 2 <sup>nd</sup> time observation allocation.....  | 14 |
| Table 4 Final time observation allocation.....            | 16 |
| Table 5 Observation distribution.....                     | 16 |
| Table 6 Class explanation.....                            | 16 |

**LIST OF FIGURES**

|  |    |
|--|----|
| Figure 1 Pair Scatterplot of Crop group.....                 | 6  |
| Figure 2 Pair Scatterplot of Livestock group.....            | 6  |
| Figure 3 Pair Scatterplot of Fishery group.....              | 7  |
| Figure 4 PCA pf crop data set .....                          | 8  |
| Figure 5 PCA of livestock data set.....                      | 9  |
| Figure 6 PCA of fish data set.....                           | 9  |
| Figure 7 Boxplot of 1 <sup>st</sup> cluster analysis.....    | 13 |
| Figure 8 Boxplot of 2 <sup>nd</sup> cluster analysis.....    | 14 |
| Figure 9 Boxplot of 10 <sup>th</sup> cluster analysis.....   | 15 |
| Figure 10 Scatter plot of cropvalue and livestockvalue ..... | 18 |
| Figure 11 Scatter plot of cropvalue and fishvalue.....       | 19 |
| Figure 12 Scatter plot of livkstockvalue and fishvalue.....  | 20 |

## 1 INTRODUCTION

Bangladesh is one of the poorest and most densely populated countries in the world, covering an area of 144,000 km<sup>2</sup> with a population of 164 million. Rice and fish have been an essential part of the life of Bangladeshi people from time immemorial. The staple foods of the people of Bangladesh are rice and fish. The demand for rice and fish is constantly increasing in Bangladesh with nearly three million people being added each year to the population of the country (Chowdhury 2009).

Crop is the largest sector of Bangladesh's agriculture. Rice is the foremost agricultural crop in Bangladesh with an annual production of over 29 million tons per annum (BRKB 2010).

The fisheries sector is a source of employment and income for a large sector of the population, particularly in rural areas. According to the fisheries statistical year book published by FRSS, this sub-sector of agriculture contributes ±5% in GDP. Fish is the source of 60% annual dietary protein of the country population demand. (Climate change and fisheries & livestock in Bangladesh)

Livestock is another integral component of agricultural economy of Bangladesh. The Bangladesh Economic Review's shows the highest growth rate of livestock sub-sector in GDP at constant prices (base year 1995-96) in the years 2004-05, 7.23% and 2005-06, 6.15% compared to crops and vegetables (4.02%) and fisheries (4.16%). Livestock is performing multifarious functions such as provisions of food, nutrition, income, saving, foreign currency earning, draft power, manure, fuel, transport, social and cultural functions. (Climate change and fisheries & livestock in Bangladesh)

In order to meet the soaring demand for food, integrated rice-fish farming plays an important role in increasing food production as the integrated farming system. It is better than rice monoculture in terms of resource utilization, diversity, productivity, and both the quality and quantity of the food produced. Nevertheless, only a small number of farmers are involved in integrated rice-fish farming due to

a lack of technical knowledge and an aversion to the risks associated with flood and drought. (Nesar and Stephen, 2011)

## **1.1 Purpose of the Study**

### **1.1.1 Background of the survey**

The Bangladesh Policy Research and Strategy Support Program (PRSSP) for Food Security and Agricultural Development, funded by the United States Agency for International Development (USAID) and implemented by the International Food Policy Research Institute (IFPRI), were launched in October 2010. To address special agricultural development issues, IFPRI researchers designed the Bangladesh Integrated Household Survey (BIHS) – the most comprehensive, nationally representative household survey conducted to date. Varied studies can use the survey’s integrated data platform to carry out research with policy implications for the country’s agricultural development. (Akhter, 2013)

The BIHS covers 6500 households in 325 primary sampling units. The sample is statistically representative at following levels: (a) nationally representative of rural Bangladesh; and (b) representative of rural areas of each of the seven administrative divisions of the country: Barisal, Chittagong, Dhaka, Khulna, Rajshahi, Rangpur, and Sylhet.

### **1.1.2 Purpose of the study**

The concept of diversification conveys different meaning to different people at different levels. In this paper, we only discuss this concept within agriculture. The diversification is considered a shift of resources from one crop (or fishery) to a larger mix of crops and fishery, and expected returns from each crop/fishery activities that leads to optimum portfolio of income.

Agricultural diversification in its standard usage, either in terms of the diversity of farm activities or markets, is a significant issue for many developing countries, such as Bangladesh, because their economics are generally characterized by the lack of it. For example, Bangladesh has traditionally relies

heavily on the rice monoculture that are predominantly vulnerable to climate variability and change.

([http://unfccc.int/adaptation/nairobi\\_work\\_programme/programme\\_activities\\_and\\_work\\_areas/items/3994.php](http://unfccc.int/adaptation/nairobi_work_programme/programme_activities_and_work_areas/items/3994.php))

For nearly 60 years, regional agricultural diversity has been promoted as a means to achieve the economic goals of stability and growth. As an economy becomes more diversified, it becomes less sensitive to fluctuations caused by factors outside of the region. (Ron, 1997) Diversification of agriculture can be used as a tool to augment farm income, generate employment, alleviate poverty and conserve precious soil and water resources. A sound understanding about the patterns of agricultural diversification and the constraints it faces would help in crafting appropriate policies regarding institutional arrangements and creation of adequate infrastructure, which could benefit a large mass of small and marginal holders. (Joshi, 2003) This study is an attempt in this direction.

This project presents results of analysis of part of the IFPRI household survey data on various topics that, combined, represents the current economic situation in Bangladesh. Specifically, the study examines the classification of agricultural activities throughout the country.

This report is organized in four sections. The first section is a basic introduction of BIHS and the purpose of the study. Section two describes the process of data cleaning and manipulation. Section 3 gives the methodology. The last section summarizes the main findings and provides the conclusions.

The result of this study intends to examine the extent, nature of economic diversification in rural Bangladesh; identify determinants of agricultural diversification.

## 2 DATA CLEANING

Three raw data sets were manipulated. The first data set, “hhmerge”, which contains the information about crop, totally has 10 variables and 6501 observations. The 10 variables are: a01, cropvalue, cropsalevalue, cropconvalue, output, outputvalue, cropirrinkindvalue, croplaborinkindvalue, price, storage. The second data set, “livestock”, containing livestock information has 8 variables and 6503 observations. The variables are: a01, sample\_type, lvkvalue, lvksalevalue, lvkconvalue, lvkcost, lvkmales, and lvkfemale. The third data set, “fishery”, containing fishery information, has 5 variables and 6502 observations. The 6 variables are: a01, fishvalue, fishsalevalue, fishconvalue, fishlabor, fishcost.

All three data sets have the common variable: a01 (household ID). Thus, the many-to-many merge technique is performed. All observations are merged based on the household level. Observations which have sample\_type =1 are dropped because they are over-sampled households. Missing values were all converted to 0. The cleaned data set contains 14 variables and 5225 observations. The variables are: a01, crop value, crop sale value, crop consumption value, livestock value, livestock sale value, livestock consumption values, livestock cost, livestock male labor, livestock female labor, fish value, fish sale value, fish consumption value, fish labor and fish cost.

### 3 CLASSIFICATION OF HOUSEHOLDS

#### 3.1 Variables grouping and PCA

##### 3.1.1 Variables grouping

Now the cleaned data set is ready to be analyzed. Obviously, this data set cannot be described only through the classical second order statistics, such as the sample mean and covariance, etc. To investigate and study the internal structure of this complex data set, cluster analysis is an important element of exploratory data analysis, which can find some unique and definitive groupings for the data.

Generally speaking, the classification of different things is a natural process for human beings. For examples, people divide fruits into separate classes by using their observable characteristics (K.Fukunaga, 1972). Tan et al. states the definition for cluster analysis from data mining point of view that "Cluster analysis divides data into groups that are meaningful, useful, or both." (P.N. Tan, M. Steinbach, and V.Kumar, 2005) By meaningful they refer to clusters that capture the natural structure of a data set, whereas the useful clusters serve only as an initial setting for some other method, such as PCA (principal component analysis) or regression methods. For these methods, it may be useful to summarize the data sets beforehand. (Sami, 2006)

Follow this definition, first of all, the natural structure of the cleaned data set is observed. Apparently, three different agricultural activities which are crop, fishery and livestock are involved in this data set. Therefore, the data set is classified into three different groups. Variables "cropvalue", "crop-salevalue" and "cropconvalue" are combined to a new data which can be called "Agriculture group"; the second "Livestock group" contains variables: "lvkvalue", "lvksalevalue", "lvkconvalue", "lvkcost", "lvkmales" and "lvkfemales"; The remaining variables "fishvalue", "fishsalevalue", "fishconvalue", "fishlabor" and "fishcost" are combined to the third new data set which can be named "Fishery group".

Next we draw the scatterplots (Figure 1, Figure 2, and Figure 3) of each new data set to detect structure in the relationships between variable.

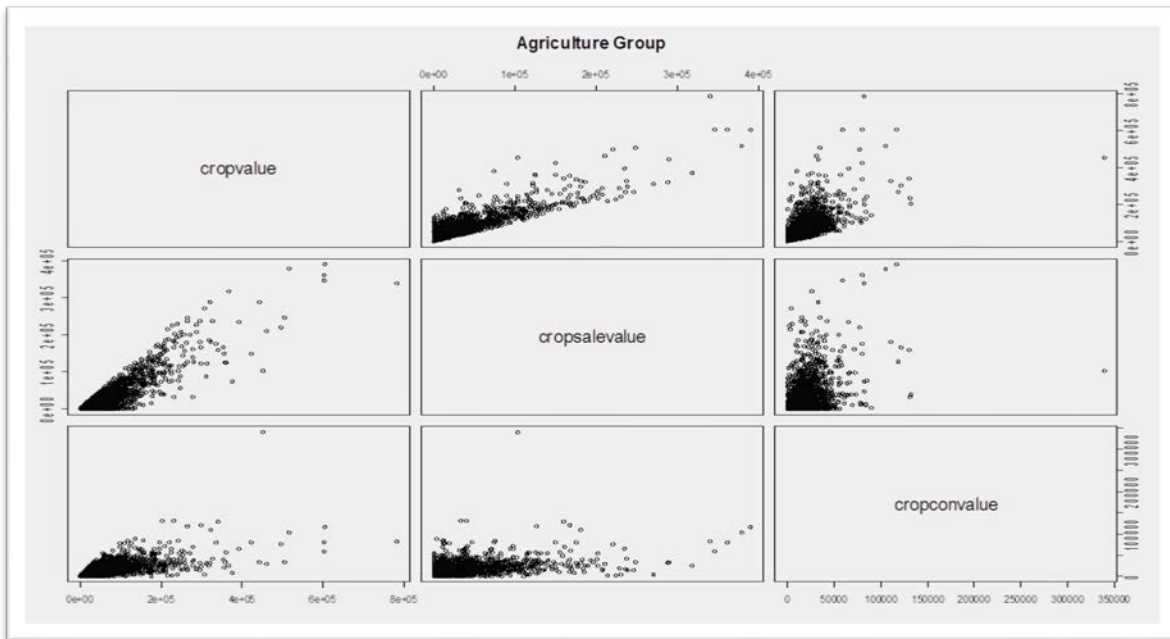


Figure 1 Pair Scatterplot of Crop Group

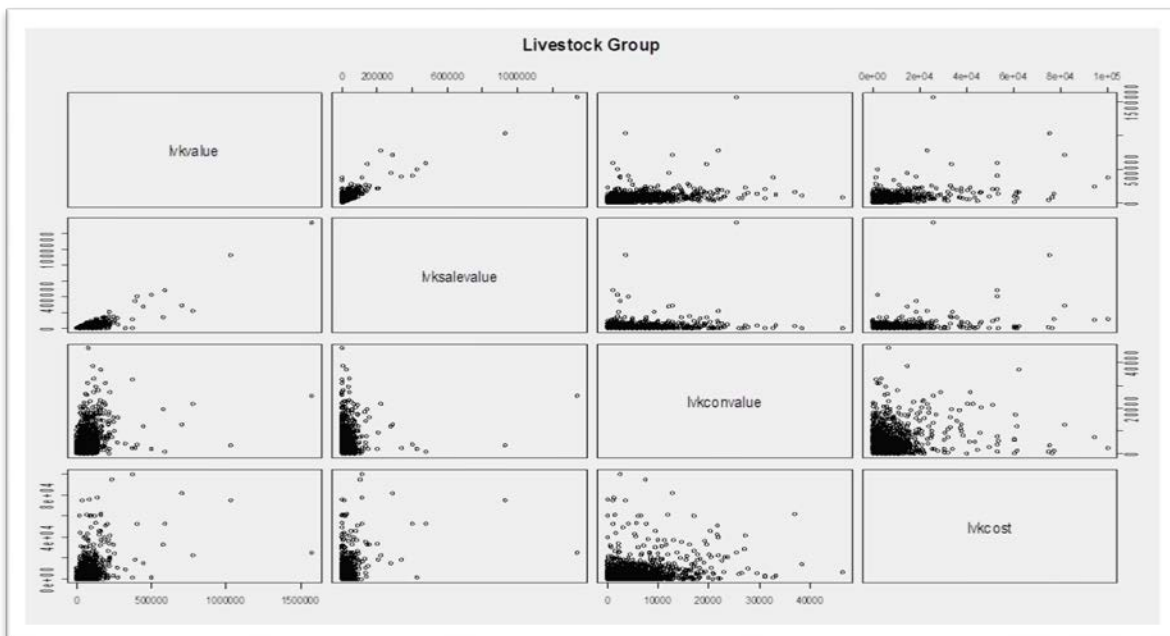


Figure 2 Pair Scatterplot of Livestock Group

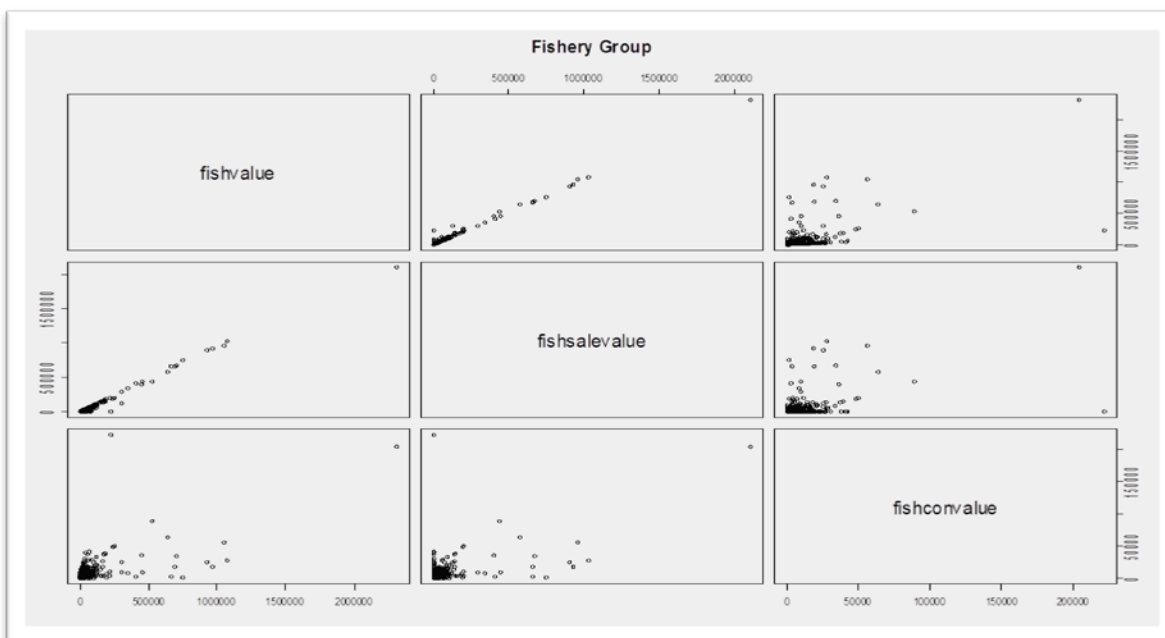


Figure 3 Pair Scatterplot of Fishery Group

### 3.1.2 *Principal component analysis*

Now PCA (Principal Component Analysis) will be performed to summarize each of three new data sets beforehand. Principal component analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables, and believe that there is some redundancy in those variables. In this study, some of the variables of each new data set are obviously correlated with one another. For example, crop value and crop sale value. In this case, it should be possible to reduce the observed variables into a smaller number of principal components that will account for most of the variance in the observed variables.

A principal component analysis was performed three times on the separated data sets. The first component of crop data set extracted in a principal component analysis accounts for 94% of total variance in the observed variables. (See Figure 4) This means that the first component will be correlated with all the other observed variables. Therefore the first component is retained, interpreted, and used in

subsequent analyses. The remaining components accounted for only 6% of variance. These latter components would therefore not be retained, interpreted, or further analyzed.

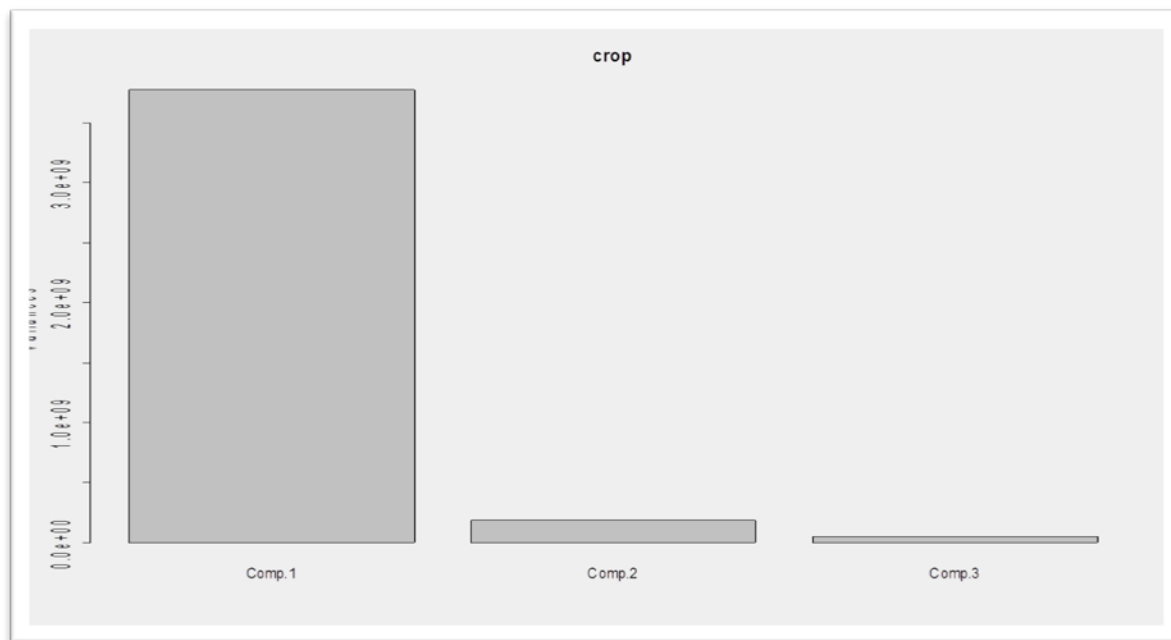


Figure 4 PCA of crop data set

Similar, The first component of livestock data set accounts for 92% of total variance. (See Figure 5) and is retained. The first component of fishery data set accounts for 94% of total variance. (See Figure 6) and is retained as well.

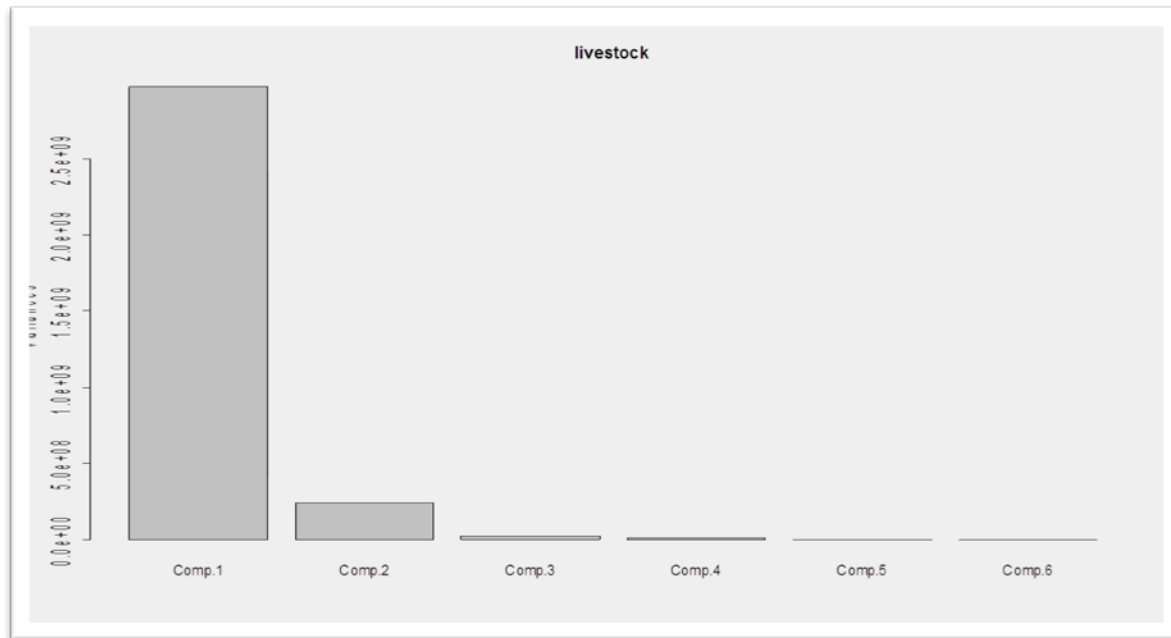


Figure 5 PCA of livestock data set

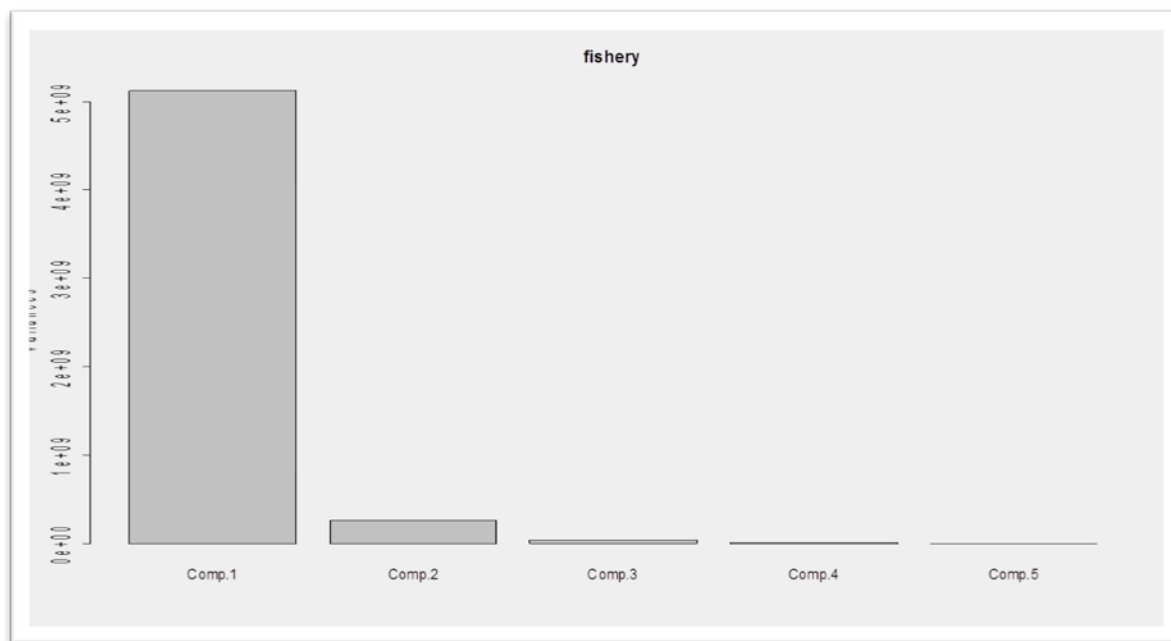


Figure 6 PCA of fishery data set

The principal component analysis was completed. Now I assigned one score to each subject to indicate that subject's standing on each retained component. With this done, these component scores could be used as criterion variables in subsequent analyses.

Next, to perform the iterative clustering analysis, three scores of the first principal component of each data set are calculated and combined to a new data set.

### **3.1.3 New data set**

The new formed data set contains 3 variables and 5225 observations. The 3 variables are: the first principal component of the crop data set; the first principal component of the fish data set and the first principal component of the livestock data set. Before data mining itself, data preprocessing plays a crucial role. Since the new data set has parameters of different units and scales, I use the standardization on my data set to transform it to have zero mean and unit variance. The equation is below:

$$X_{new} = \frac{X - \mu}{\sigma}$$

Till now, the transformed data set is ready to be investigated and analyzed. Obviously, our data set can not be described only through the classical second order statistics, which include sample mean and covariance. To obtain qualitative and quantitative understanding of this large amount of multivariate data set and find its internal structure, cluster analysis is an important data mining technique. We will describe this method and the applications in the next chapter.

## **3.2 Iterative clustering**

Data clustering is an important data mining technique, which offers refined and more abstract views to the inherent structure of a data set by partitioning it into a number of disjoint or overlapping groups. This study will perform a special class of clustering algorithms, namely partition-based methods. These clustering methods are flexible methods based on iterative relocation of data points between clusters. The quality of the solutions is measured by a clustering criterion. At each iteration, the iterative relocation algorithms reduce the value of the criterion function until convergence. (Sami, 2006)

The upper and lower quartile values of a data set are always be chosen as the criterion to divide the data set. The definition is: A pth percentile value is a number which puts at least P percent of the data

values at that number or below and at least (100-P) percent of the data values at that number or above. (<http://www.amstat.org/publications/jse/v14n3/langford.html>) In this project, the 1<sup>st</sup> quartile value and the 3<sup>rd</sup> quartile value of the first and second columns of the new data set are chosen to be the criterion to divide the data set. However, for the third column of the new data set, I only choose the 3<sup>rd</sup> quartile as the criterion because only 64% household have fish value records. Below is the table to explain how I assign the observations to different classes.

Table 1: Clustering criterion

|         | 1 <sup>st</sup> PC of Crop data set | 1 <sup>st</sup> PC of Livestock data set | 1 <sup>st</sup> PC of Fishery data set |
|---------|-------------------------------------|--|--|
| Class 1 | $\geq 75\% Y_{[1]}$                 | $< 25\% Y_{[2]}$                         | $< 75\% Y_{[3]}$                       |
| Class 2 | $< 25\% Y_{[1]}$                    | $\geq 75\% Y_{[2]}$                      | $< 75\% Y_{[3]}$                       |
| Class 3 | $< 25\% Y_{[1]}$                    | $< 25\% Y_{[2]}$                         | $\geq 75\% Y_{[3]}$                    |
| Class 4 | $\geq 75\% Y_{[1]}$                 | $\geq 75\% Y_{[2]}$                      | $< 75\% Y_{[3]}$                       |
| Class 5 | $\geq 75\% Y_{[1]}$                 | $< 25\% Y_{[2]}$                         | $\geq 75\% Y_{[3]}$                    |
| Class 6 | $< 25\% Y_{[1]}$                    | $\geq 75\% Y_{[2]}$                      | $\geq 75\% Y_{[3]}$                    |
| Class 7 | $\geq 75\% Y_{[1]}$                 | $\geq 75\% Y_{[2]}$                      | $\geq 75\% Y_{[3]}$                    |

To perform the algorithm, I first set 7 null classes. The observations whose 1<sup>st</sup> column values are greater than the 3<sup>rd</sup> quartile, meanwhile, the second column values are smaller than the 1<sup>st</sup> quartile, and third column values are smaller than the 3<sup>rd</sup> quartile are assigned to the first class. Therefore class 1 is defined as “Crop class” for the present. Similarly, class 2, “Livestock class”, contains households who have largest livestock value while crop and fishery values are smallest. Class 3 is defined as “Fishery class”. Class 4 is for households who have large crop and livestock values and small fish values. Class 5 is for households who have large crop and fish values and small livestock values. Class 6 is for households who have large livestock and fish values small crop values. The last class contains households who have large crop, livestock and fish values.

To discover structures in data, the iterative cluster analysis is performed. The cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. (<http://www.statsoft.com/textbook/cluster-analysis/>) Clustering plays an important role in a large number of examples. In this project, the clustering method is performed to classify values into groups, and we need to find out the factors which determine the difference of groups and how the factors influence the structure of Bangladesh's rural economics.

Next, the first time clustering analysis is applied. (See Figure 7) From the figure 7, we can easily observe that cluster 3,5,8,9 contains the greatest crop values or livestock values or fish values. The observations with extreme values are extracted. By the criteria I just set, some observations are allocated to the seven classes. The other observations which do not meet the criterion are not chosen.

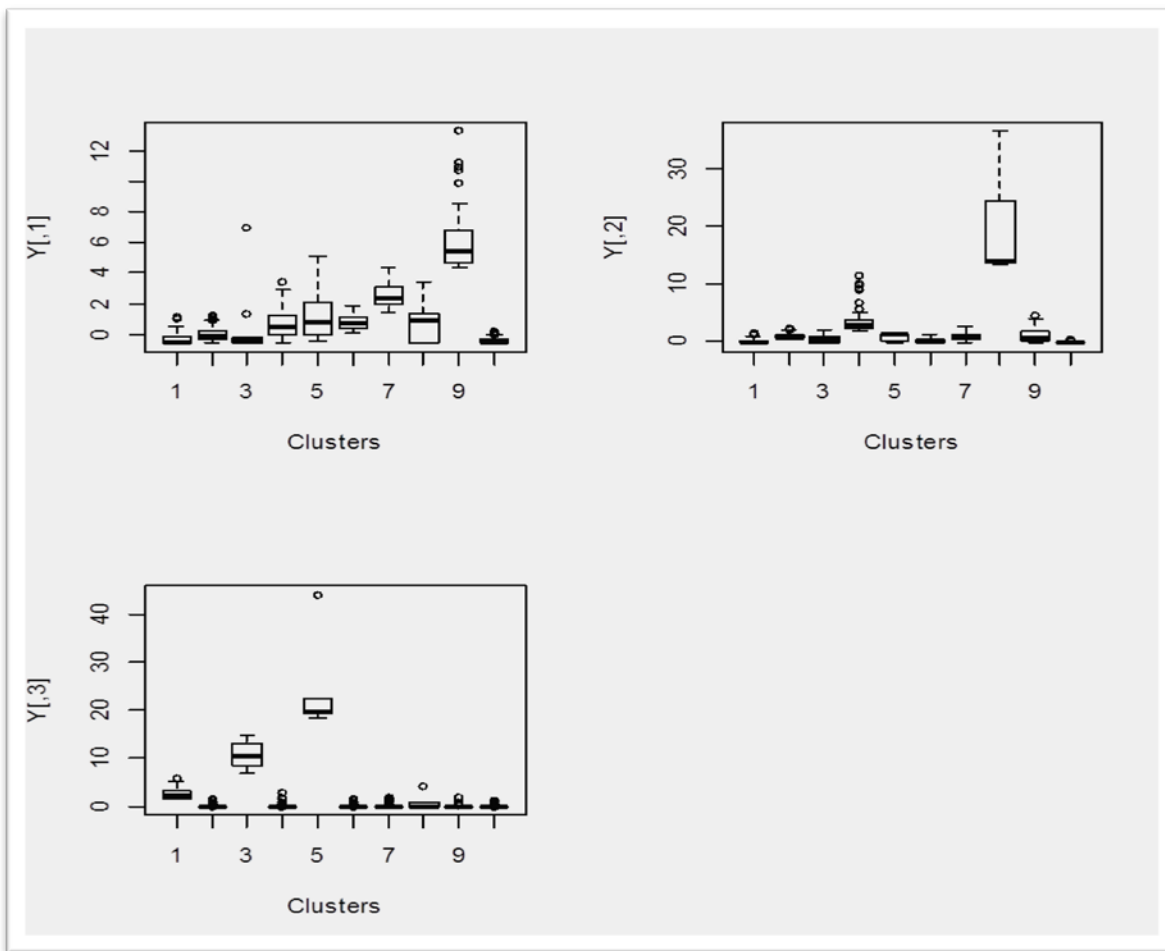


Figure 7 Boxplot of 1<sup>st</sup> cluster analysis

Table 2 1<sup>st</sup> time observations allocation

| Class.1 | Class.2 | Class.3 | Class.4 | Class.5 | Class.6 | Class.7 |
|---------|---------|---------|---------|---------|---------|---------|
| 1       | 2       | 0       | 19      | 0       | 1       | 23      |

Totally, 46 outstanding observations were chosen and removed from the original data set. The new data set again was divided to the three groups: Crop, Livestock and Fishery. Followed by the PCA, the first principal components of each group were combined to a new data set. Standardization was applied to the new data set, and the second cluster analysis was ready to be performed exactly the same as the first cluster analysis.

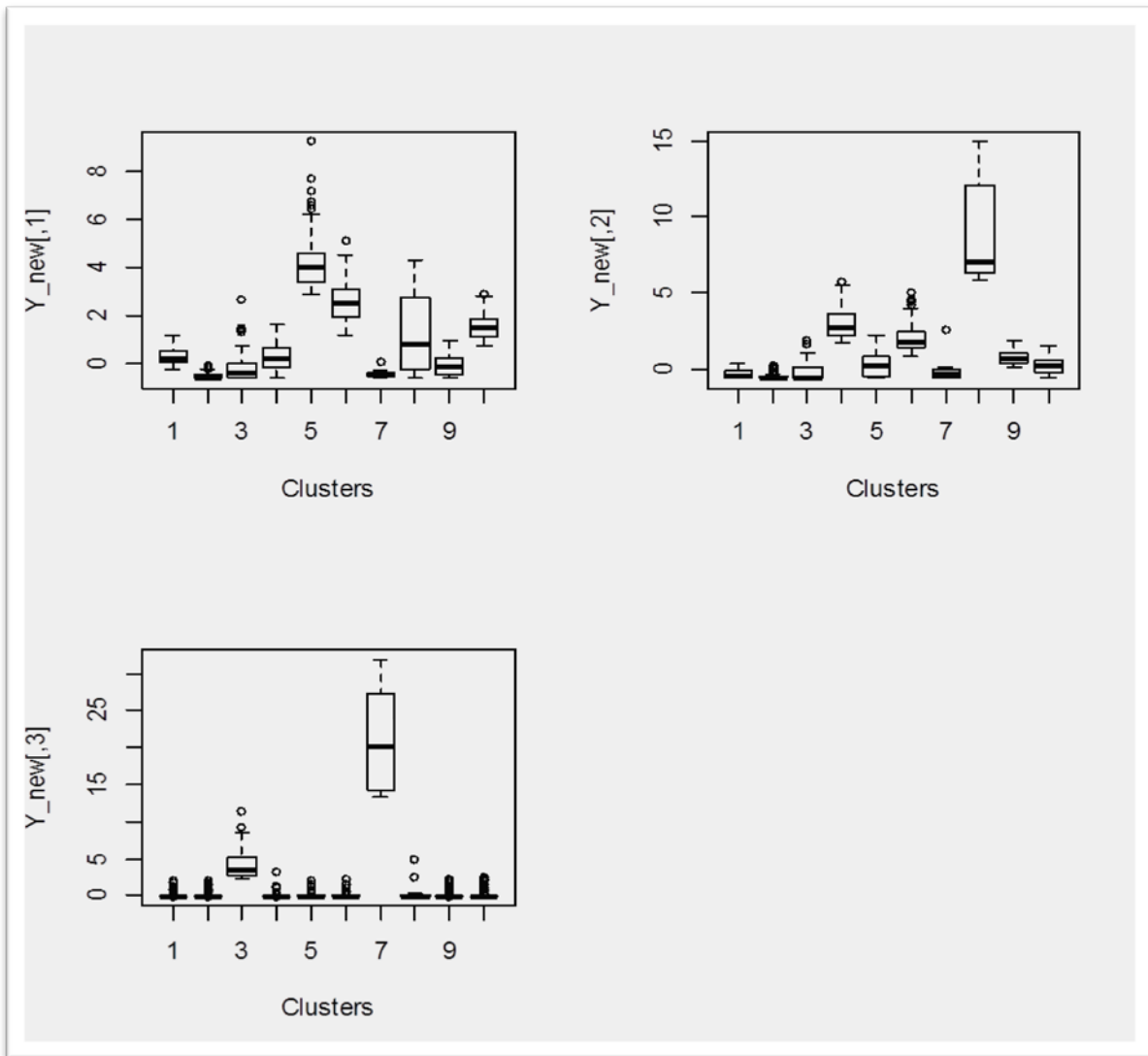


Figure 8 Boxplot of 2<sup>nd</sup> cluster analysis

Table 3 2<sup>nd</sup> time observations allocation

| Class.1 | Class.2 | Class.3 | Class.4 | Class.5 | Class.6 | Class.7 |
|---------|---------|---------|---------|---------|---------|---------|
| 5       | 3       | 4       | 48      | 1       | 4       | 61      |

After the second cluster analysis, totally 126 observations are removed from the original data

set.

The same procedure was carried on totally ten times. Figure 9 is the boxplot of the last cluster analysis.

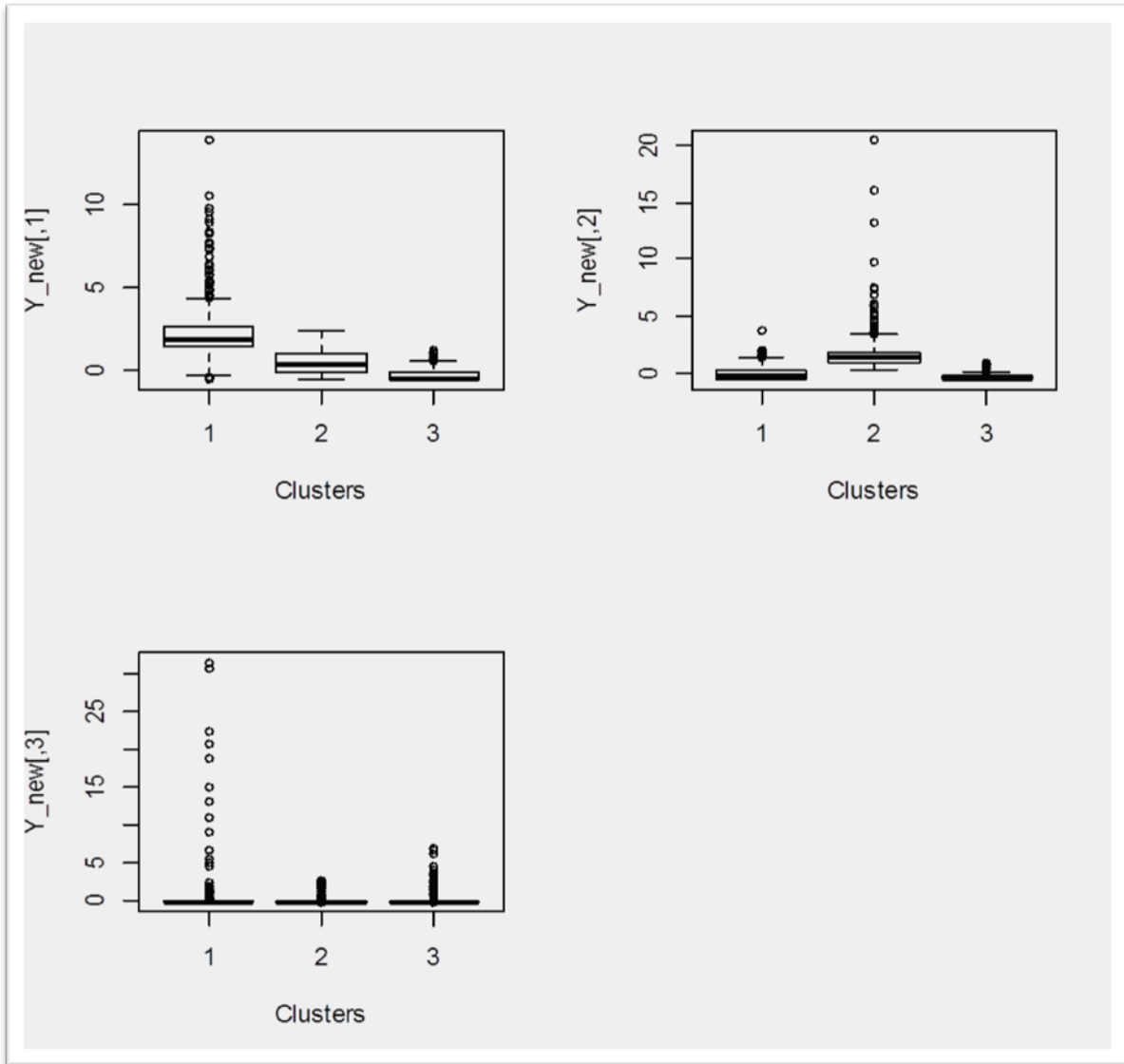


Figure 9 Boxplot of 10<sup>th</sup> cluster analysis

Obviously there were no outstanding observations after ten times cluster analysis. The left observations were all not meeting the criterion. Therefore I assigned the remaining observations into three new classes. (See Table 4)

Table 4 Final time observations allocation

| Class.1 | Class.2 | Class.3 | Class.4 | Class.5 | Class.6 | Class.7 | Class.8 | Class.9 | Class.10 |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| 28      | 8       | 6       | 251     | 17      | 7       | 214     | 379     | 866     | 3449     |

### 3.3 Result

The iterative clustering analysis is performed totally 10 times. 5225 observations are finally assigned to 10 classes reordered by the frequency. (See Table 5)

Table 5 Observations distribution

| Class.1 | Class.2 | Class.3 | Class.4 | Class.5 | Class.6 | Class.7 | Class.8 | Class.9 | Class.10 |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|----------|
| 3449    | 866     | 379     | 251     | 214     | 28      | 17      | 8       | 7       | 6        |

Now the classes have different meaning. Class 1, 2 and 3 contains households who are poor farmers neither producing agriculture values nor livestock or fishery values. Class 4 contains households who are doing integrated crop-livestock farming. Class 5 presents the richest farmers who produce most agriculture, livestock and fishery values. Class 6 has the households who are only doing agriculture. Class 7 contains households who are good at integrated crop-fish farming. Class 8 has the farmers who are only raising livestock. Class 9 contains the households who are doing integrated livestock-fishery farming, but seldom doing agriculture. The last class has the households who are only doing fishery. (See table 6)

Table 6 Class explanation

| Class Number | Crop Value | Livestock Value | Fish Value |
|--------------|------------|-----------------|------------|
| 1            | Very Low   | Very Low        | Very Low   |
| 2            | Low        | Very Low        | Very Low   |
| 3            | Low        | Very Low        | Low        |
| 4            | High       | High            | Low        |
| 5            | High       | High            | High       |
| 6            | High       | Low             | Low        |
| 7            | High       | Low             | High       |
| 8            | Low        | High            | Low        |

|    |     |      |      |
|----|-----|------|------|
| 9  | Low | High | High |
| 10 | Low | Low  | High |

To get a clear picture of how the total households are distributed in 10 classes. The scatter plots are draw below. (See Figure 10, Figure 11, Figure 12)

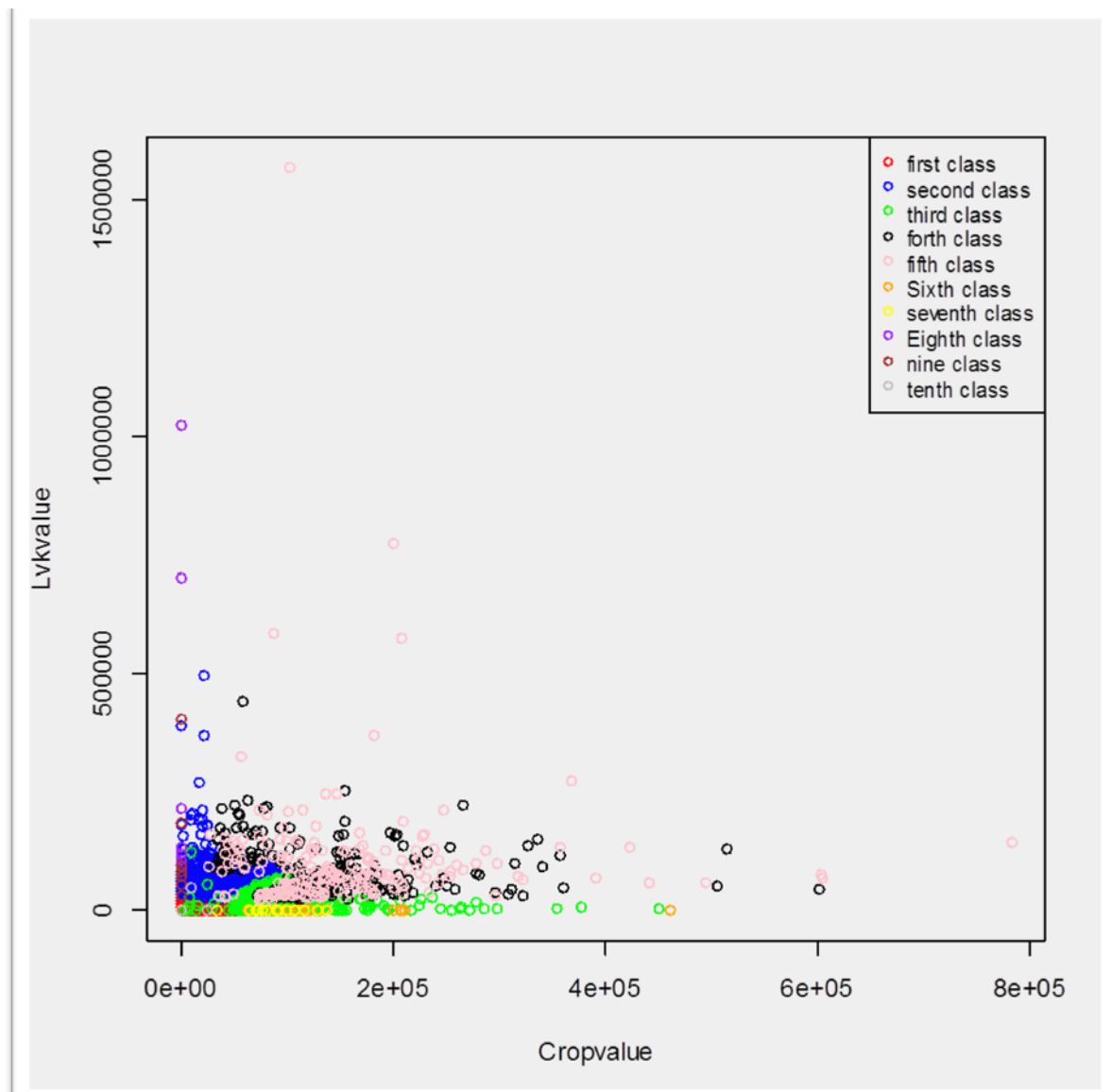


Figure 10 Scatter plot of cropvalue and livestockvalue

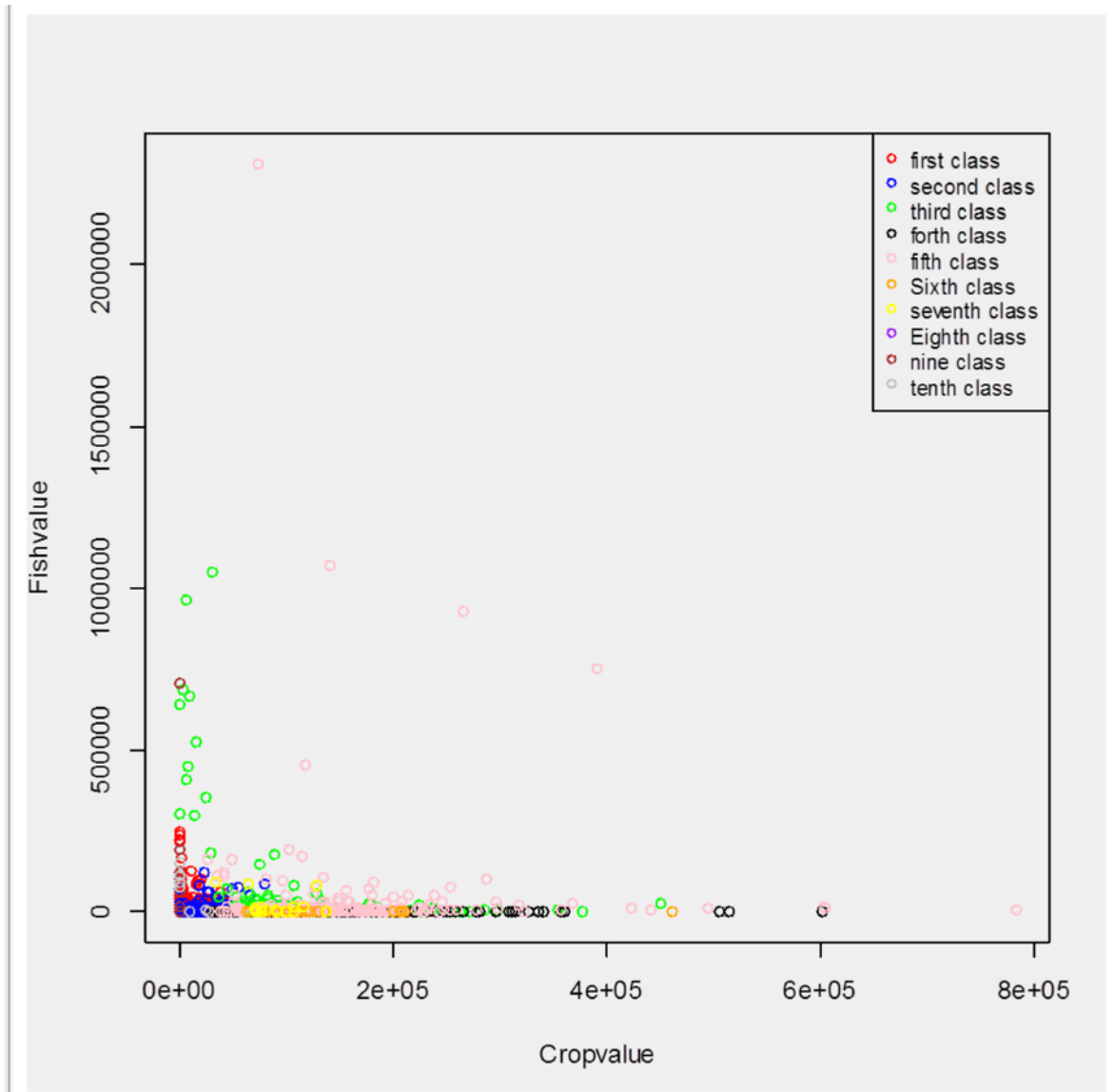


Figure 11 Scatter plot of cropvalue and fishvalue

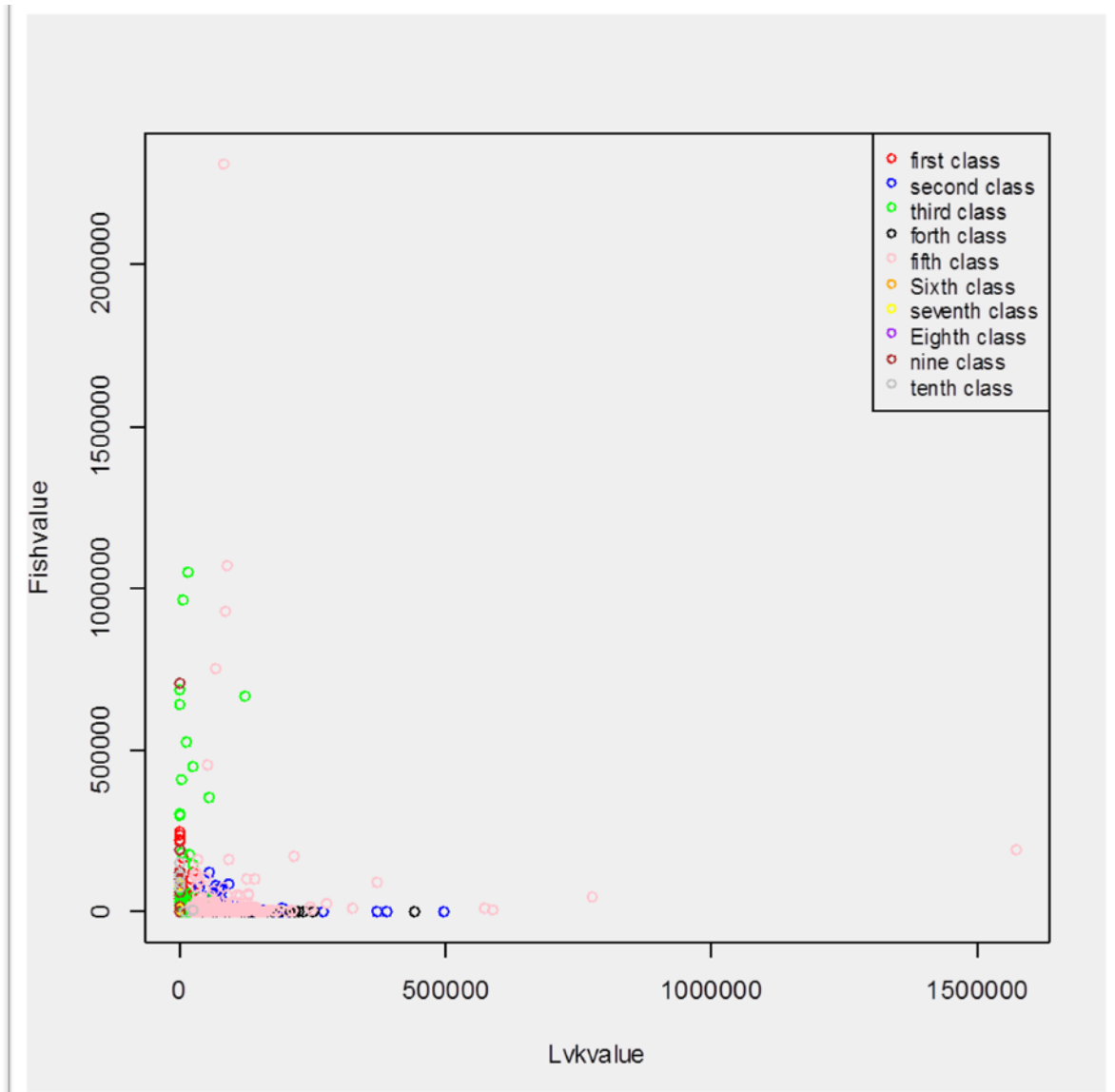


Figure 12 Scatter plot of livestockvalue and fishvalue

#### 4 CONCLUSIONS

The thorough understanding of the factors influencing the agricultural diversification of rural Bangladesh is crucial for effective policy design to improve the food security of the Bangladeshi people. The iterative clustering analysis technique has been reviewed in this paper. This approach solve the problem of categorizing data by partitioning a data set into a number of clusters based on some similarity measure so that the similarity in each cluster is larger than among clusters. This method has been implemented and tested against Bangladesh integrated survey data.

This study reveals that crop-fish, livestock-fish, crop-livestock integration could be a visible option for diversification. Such farm diversification will enhance food security, and should play an important role in contributing to food security in Bangladesh. However, the integrated farming has not yet been attempted on a large scale in the country. Therefore, the integrated farming types should be encouraged and helped by the government. It is necessary to provide institutional and organizational support, training facilities and technical support for sustainable integrated farming. Training and technical support would help to increase the knowledge of farmers, improve productivity and reduce risks. In addition, further research would be required on social, economic, environmental, and livelihood issues for the adoption of integrate farming in rural Bangladesh. (Nesar and Stephen, 2011)

## REFERENCES

- [1] Sami Ayramo and Tommi Karkkainen, *Introduction to partitioning-based clustering methods with a robust example*, Report of the Department of Mathematical Information Technology Series C. Software and Computational Engineering, No. C. 1/2006
- [2] P. K. Joshi, Ashok Gulati, Pratap S. Birthal, and Laxmi Tewari, *Agriculture Diversification in South Asia: Patterns, Determinants, and Policy Implications*, Mssd Discussion Paper NO.57
- [3] Nesar Ahmed and Stephen T. Garnett, *Integrated rice-fish farming in Bangladesh: meeting the challenges of food security*, Springer Science + Business Media B. V. & International Society for Plant Pathology 2011
- [4] Khaled Hammouda, *A comparative Study of Data Clustering Techniques*
- [5] Frank Lin and William W. Cohen, *Power Iteration Clustering*, Appearing in Proceedings of the 27<sup>th</sup> International Conference on Machine Learning, Haifa, Israel, 2010
- [5] Kim, J. O. & Mueller, C. W. (1978a), *Introduction to factor analysis: What it is and how to do it*, Beverly Hills, CA: Sage
- [6] Kim, J. O. & Mueller, C. W. (1978b), *Factor analysis: Statistical methods and practical issues*, Beverly Hills, CA: Sage
- [7] Stevens, J. (1986), *Applied Multivariate Statistics for the Social Sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates
- [8] K.Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Inc, 1972
- [9] P.N. Tan, M. Steinbach, and V.Kumar, *Introduction to data mining*, Addison-Wesley, 2005
- [10] Ron Shaffer, *The role of Diversification in Economic Growth and Stability*, 2007

## APPENDIX: R SCRIPT

```
#set 8 null classes

class.1=NULL #crop

class.2=NULL #livestock

class.3=NULL #fish

class.4=NULL #crop livestock

class.5=NULL #crop fish

class.6=NULL #livestock fish

class.7=NULL #crop livestock fish large

class.8=NULL #crop livestock fish small

#read the cleaned data

hh=read.csv("hhh.csv")

hh[is.na(hh)] <- 0 #convert NA to 0

#non-agriculture household

D<-hh[!(hh$agvalue==0),]

X=as.matrix(D[,-c(1,16,17,18,19)])

#Pairs scatterplot

pairs(~cropvalue+cropsalevalue+cropconvalue,data=crop,

      main="Agriculture Group")

pairs(~lvkvalue+lvksalevalue+lvkconvalue+lvkcost+lvkmale+lvkfemale,data=lvk,

      main="Livestock Group")
```

```
pairs(~fishvalue+fishsalevalue+fishconvalue,data=fish,  
      main="Fishery Group")
```

```
#divide the data set to three categories
```

```
crop=X[,c(1,2,3)] #crop household
```

```
lvk=X[,c(4:9)] #lvk househod
```

```
fish=X[,c(10:14)] #fishery houseld
```

```
#PCA
```

```
source("spcaBP.R")
```

```
lambda=0
```

```
PCs1=spca.BP(crop,1,lambda)
```

```
Y1=crop%*%PCs1
```

```
PCs2=spca.BP(lvk,1,lambda)
```

```
Y2=lvk%*%PCs2
```

```
PCs3=spca.BP(fish,1,lambda)
```

```
Y3=fish%*%PCs3
```

```
#combine three scores and scale
```

```
Y=cbind(Y1,Y2,Y3)
```

```
Y=scale(Y)
```

```
row.names(Y)=NULL
```

```
crop_25=as.numeric(quantile(Y[,1],.25))
```

```
crop_75=as.numeric(quantile(Y[,1],.75))
```

```

lvk_25=as.numeric(quantile(Y[,2],.25))
lvk_75=as.numeric(quantile(Y[,2],.75))
fish_25=as.numeric(quantile(Y[,3],.25))
fish_75=as.numeric(quantile(Y[,3],.75))

#cluster analysis

clu.fit= kmeans(Y, 10)

str(clu.fit)

#boxplot

par(mfrow=c(2,2))

boxplot(Y[,1]~clu.fit$cluster,xlab="Clusters",ylab="Y[,1]")
boxplot(Y[,2]~clu.fit$cluster,xlab="Clusters",ylab="Y[,2]")
boxplot(Y[,3]~clu.fit$cluster,xlab="Clusters",ylab="Y[,3]")

#delete observations

temp1= which(clu.fit$cluster%in%c(3,5,8,9))

class.1=temp1[which(Y[temp1,1] >= crop_75 & Y[temp1,2] < lvk_25 & Y[temp1,3] < fish_75)]
class.2=temp1[which(Y[temp1,1] < crop_25 & Y[temp1,2] >= lvk_75 & Y[temp1,3] < fish_75 )]
class.3=temp1[which(Y[temp1,1] < crop_25 & Y[temp1,2] < lvk_25 & Y[temp1,3] >= fish_75)]
class.4=temp1[which(Y[temp1,1] >= crop_75 & Y[temp1,2] >= lvk_75 & Y[temp1,3] < fish_75)]
class.5=temp1[which(Y[temp1,1] >= crop_75 & Y[temp1,2] < lvk_25 & Y[temp1,3] >= fish_75)]
class.6=temp1[which(Y[temp1,1] < crop_25 & Y[temp1,2] >= lvk_75 & Y[temp1,3] >= fish_75)]

```

```
class.7=temp1[which(Y[temp1,1] >= crop_75 & Y[temp1,2] >= lvk_75 & Y[temp1,3] >= fish_75)]
```

```
class.8=temp1[which(Y[temp1,1] < crop_25 & Y[temp1,2] < lvk_25 & Y[temp1,3] < fish_75)]
```

```
remove=c(class.1,class.2,class.3,class.4,class.5,class.6,class.7,class.8)
```

```
X_new=X[-remove,]
```

```
left=(1:dim(X)[1])[-remove]
```

```
#cluster analysis
```

```
crop_new=X_new[,c(1,2,3)] #crop household
```

```
lvk_new=X_new[,c(4:9)] #lvk househod
```

```
fish_new=X_new[,c(10:14)] #fishery houseld
```

```
PCs1=spca.BP(crop_new,1,lambda)
```

```
Y1=crop_new%%PCs1
```

```
PCs2=spca.BP(lvk_new,1,lambda)
```

```
Y2=lvk_new%%PCs2
```

```
PCs3=spca.BP(fish_new,1,lambda)
```

```
Y3=fish_new%%PCs3
```

```
Y1=Y1*sign(Y1[which(abs(Y1)==max(abs(Y1)))])
```

```
Y2=Y2*sign(Y2[which(abs(Y2)==max(abs(Y2)))])
```

```
Y3=Y3*sign(Y3[which(abs(Y3)==max(abs(Y3)))])
```

```
Y_new=cbind(Y1,Y2,Y3)
```

```
Y_new=scale(Y_new)
row.names(Y_new)=NULL

crop_25=quantile(Y_new[,1],.25)
crop_75=quantile(Y_new[,1],.75)
lvk_25=quantile(Y_new[,2],.25)
lvk_75=quantile(Y_new[,2],.75)
fish_25=quantile(Y_new[,3],.25)
fish_75=quantile(Y_new[,3],.75)

fit= kmeans(Y_new, 3)
str(fit)

par(mfrow=c(2,2))
boxplot(Y_new[,1]~fit$cluster,xlab="Clusters",ylab="Y_new[,1]")
boxplot(Y_new[,2]~fit$cluster,xlab="Clusters",ylab="Y_new[,2]")
boxplot(Y_new[,3]~fit$cluster,xlab="Clusters",ylab="Y_new[,3]")

temp1=which(fit$cluster%in%c(1,5,9))

class.1=c(class.1,left[temp1[which(Y_new[temp1,1] >= crop_75 & Y_new[temp1,2] < lvk_25 &
Y_new[temp1,3] < fish_75 )]])
class.2=c(class.2,left[temp1[which(Y_new[temp1,1] < crop_25 & Y_new[temp1,2] >= lvk_75 &
Y_new[temp1,3] < fish_75 )]])
```

```

class.3=c(class.3,left[temp1[which(Y_new[temp1,1] < crop_25 & Y_new[temp1,2] < lvk_25 &
Y_new[temp1,3] >= fish_75))])
class.4=c(class.4,left[temp1[which(Y_new[temp1,1] >= crop_75 & Y_new[temp1,2] >= lvk_75 &
Y_new[temp1,3] < fish_75))])
class.5=c(class.5,left[temp1[which(Y_new[temp1,1] >= crop_75 & Y_new[temp1,2] < lvk_25 &
Y_new[temp1,3] >= fish_75))])
class.6=c(class.6,left[temp1[which(Y_new[temp1,1] < crop_25 & Y_new[temp1,2] >= lvk_75 &
Y_new[temp1,3] >= fish_75))])
class.7=c(class.7,left[temp1[which(Y_new[temp1,1] >= crop_75 & Y_new[temp1,2] >= lvk_75 &
Y_new[temp1,3] >= fish_75))])
class.8=c(class.8,left[temp1[which(Y_new[temp1,1] < crop_25 & Y_new[temp1,2] < lvk_25 &
Y_new[temp1,3] < fish_75))])

par(mfrow=c(2,2))
plot(X[left,1],X[left,4],xlab="Cropvalue",ylab="Lvkvalue")
plot(X[left,1],X[left,10],xlab="Cropvalue",ylab="fishvalue")
plot(X[left,4],X[left,10],xlab="Lvkvalue",ylab="fishvalue")
class3=left[which(fit$cluster==1)]# crop+fish larger #379
class2=left[which(fit$cluster==2)]#larger lvkvalue #866
class1=left[which(fit$cluster==3)]# crop+fish smaller #3449
class4=class.4 #crop lvk 251
class5=class.7 #crop lvk fish 214
class6=class.1 #crop 28
class7=class.5 #crop fish 17

```

```
class8=class.2 #lvk 8
```

```
class9=class.6 #lvk fish 7
```

```
class10=class.3 #fish 6
```

```
class.1=class1
```

```
class.2=class2
```

```
class.3=class3
```

```
class.4=class4
```

```
class.5=class5
```

```
class.6=class6
```

```
class.7=class7
```

```
class.8=class8
```

```
class.9=class9
```

```
class.10=class10
```

```
plot(X[,1],X[,4],xlab="Cropvalue",ylab="Lvkvalue",type="n")
```

```
points(X[class.1,1],X[class.1,4],col="red")
```

```
points(X[class.2,1],X[class.2,4],col="blue")
```

```
points(X[class.3,1],X[class.3,4],col="green")
```

```
points(X[class.4,1],X[class.4,4],col="black")
```

```
points(X[class.5,1],X[class.5,4],col="pink")
```

```
points(X[class.6,1],X[class.6,4],col="orange")
```

```
points(X[class.7,1],X[class.7,4],col="yellow")
```

```
points(X[class.8,1],X[class.8,4],col="purple")
```

```
points(X[class.9,1],X[class.9,4],col="brown")
```

```

points(X[class.10,1],X[class.10,4],col="grey")

legend("topright", c("first class","second class","third class","forth class","fifth class", "Sixth
class","seventh class","Eighth class","nine class","tenth class"), cex=0.8,
col=c("red","blue","green","black","pink","orange","yellow","purple","brown","grey"), pch=1)

```

```

plot(X[,1],X[,10],xlab="Cropvalue",ylab="Fishvalue",type="n")

```

```

points(X[class.1,1],X[class.1,10],col="red")

```

```

points(X[class.2,1],X[class.2,10],col="blue")

```

```

points(X[class.3,1],X[class.3,10],col="green")

```

```

points(X[class.4,1],X[class.4,10],col="black")

```

```

points(X[class.5,1],X[class.5,10],col="pink")

```

```

points(X[class.6,1],X[class.6,10],col="orange")

```

```

points(X[class.7,1],X[class.7,10],col="yellow")

```

```

points(X[class.8,1],X[class.8,10],col="purple")

```

```

points(X[class.9,1],X[class.9,10],col="brown")

```

```

points(X[class.10,1],X[class.10,10],col="grey")

```

```

legend("topright", c("first class","second class","third class","forth class","fifth class", "Sixth
class","seventh class","Eighth class","nine class","tenth class"), cex=0.8,

```

```

col=c("red","blue","green","black","pink","orange","yellow","purple","brown","grey"), pch=1)

```

```

plot(X[,4],X[,10],xlab="Lvkvalue",ylab="Fishvalue",type="n")

```

```

points(X[class.1,4],X[class.1,10],col="red")

```

```

points(X[class.2,4],X[class.2,10],col="blue")

```

```

points(X[class.3,4],X[class.3,10],col="green")

```

```
points(X[class.4,4],X[class.4,10],col="black")
points(X[class.5,4],X[class.5,10],col="pink")
points(X[class.6,4],X[class.6,10],col="orange")
points(X[class.7,4],X[class.7,10],col="yellow")
points(X[class.8,1],X[class.8,10],col="purple")
points(X[class.9,1],X[class.9,10],col="brown")
points(X[class.10,1],X[class.10,10],col="grey")

legend("topright", c("first class","second class","third class","forth class","fifth class", "Sixth
class","seventh class","Eighth class","nine class","tenth class"), cex=0.8,
col=c("red","blue","green","black","pink","orange","yellow","purple","brown","grey"), pch=1)

save(class.1,file="class.1.RData")
save(class.2,file="class.2.RData")
save(class.3,file="class.3.RData")
save(class.4,file="class.4.RData")
save(class.5,file="class.5.RData")
save(class.6,file="class.6.RData")
save(class.7,file="class.7.RData")
save(class.8,file="class.8.RData")
save(class.9,file="class.9.RData")
save(class.10,file="class.10.RData")
```