Using Association Rules and Logical Learning for Clustering

by

Jia Liu

A Thesis
presented to
The University of Guelph

In partial fulfillment of requirements
for the degree of
Master of Science
in
School of Computer Science

Guelph, Ontario, Canada

© Jia Liu, August, 2014
Using Association Rules and Logical Learning for Clustering

Jia Liu
University of Guelph, 2014

Advisor:
Dr. Fei Song
Dr. David Swayne

This thesis tests whether the apriori algorithm, together with Current Best Hypothesis logical learning, can cluster data such as environmental data. The procedure is tested on a large, sparse dataset describing the ecological impact of agriculture on macroinvertebrate populations in the TaiZi River Basin of the Peoples Republic of China. We compare the association rules / CBH results with two other established algorithms. We demonstrate that our results are a considerable improvement over the other methods tested. We confirm the hypothesis that a threshold of roughly 21% or more of land area devoted to agriculture in this watershed is the tipping point to surface water degradation. So-called stochastic dominance is used to further understand the utility of agricultural land use and mitigation procedures such as riparian barriers to intercept and treat runoff. The new algorithm has been successful, but more study of the various apriori parameters is necessary.
Acknowledgments

First, I would like to thank to my advisor Dr. Fei Song and co-advisor Dr. David Swayne for the support of my MSc. study and research. They are very enthusiastic and knowledgeable, especially Dr. Swayne, as he guided me all the time during the research and helped me with the writing of the thesis.

I would also like to thank my committee member Prof. Charlie Obimbo who helped me to review my progress report.

Finally, I would like to thank my parents Wei Liu and Changjin Feng for giving birth to me, and all further support during my life.
Contents

List of Tables vi
List of Figures viii

1 Introduction
1.1 Problem Statement .......................................................... 1
1.2 Thesis Statement ............................................................... 2
1.3 Objectives ........................................................................ 3
1.4 Research Area and Sampling Locations ................................. 4
1.5 Overview of Thesis ............................................................ 6

2 Background and Literature Review
2.1 Data Mining and Cluster Analysis ......................................... 7
2.2 Discussion of Association Rules .............................................. 9
  2.2.1 Basket Analysis .......................................................... 9
  2.2.2 Basic Description of Association Rules .............................. 9
  2.2.3 Decomposition of Association Rules ................................. 11
2.3 Apriori Algorithm .............................................................. 12
2.4 K-means Algorithm ............................................................ 15
2.5 Knowledge Learning and Current Best Hypothesis Search Algorithm . 17
2.6 Convex Hull and Graham Scan Algorithm ............................ 21
2.7 Stochastic Dominance ........................................................ 24

3 Methodology and Implementation
3.1 Dataset Management .......................................................... 27
  3.1.1 Scientific and Standardized Data Structure as the Basis of Apriori 27
  3.1.2 Suitability of the Problem for this Study ......................... 28
  3.1.3 A Reasonable Division of the Data Range ....................... 29
  3.1.4 Data Standardization .................................................... 31
  3.1.5 Items Group and Logical Operators ................................ 33
  3.1.6 Defining the Output File Name Format ........................... 33
3.2 Creation of Rules .............................................................. 34
3.3 Training and Testing Rules .................................................. 36
  3.3.1 Sorting the Items ........................................................ 36
  3.3.2 Effect of Sorting the Training Data ............................... 37
  3.3.3 Training Parameter Setting ......................................... 38
4 Evaluation

4.1 Landuse Data Statistics ................................. 42
4.2 Analysis of Apriori Results ............................... 43
4.3 Current Best Hypothesis Analysis ......................... 45
    4.3.1 Sorting Method ...................................... 45
    4.3.2 Sampling Sites Cluster Type ......................... 48
    4.3.3 Training Results & Test Matching Rate ............... 51
4.4 Reduction in Numbers of Sampling Sites for Training .... 53
4.5 Dominance Analysis ....................................... 53
4.6 Current Best Hypothesis vs. Beijing Results .............. 56
    4.6.1 Number of Sites Disagreeing ....................... 56
4.7 Two Dimensional Plot ..................................... 63
    4.7.1 Box Plot .............................................. 64
    4.7.2 Outliers Analysis ................................... 67
4.8 Current Best Hypothesis vs. K-means ...................... 71
4.9 Differences Between Three Algorithms .................... 73

5 Conclusions and Future Work ............................... 74

5.1 Conclusions ............................................. 74
5.2 Future work ............................................. 77

Bibliography .................................................. 78

A Training Results Files .................................. 81
B Training Results Contents ............................... 82
C CBH Code Generation .................................... 85
List of Tables

1.1 Original Species Minus <1% Rare Species ........................................... 6
2.1 Example Taizi River Database with Transactions ................................. 14
2.2 Example: Support ............................................................................. 14
2.3 Example: Support (Continued) ......................................................... 14
2.4 Example: Support (Continued) ......................................................... 15
3.1 Ranges & Number of Rules ............................................................... 30
3.2 Data Range Parameter Setting .......................................................... 30
3.3 Data Range ...................................................................................... 31
3.4 Format of Apriori Result .................................................................. 32
3.5 Transactions in R ............................................................................ 34
3.6 Transaction ID and Items ................................................................. 34
3.7 Apriori in R .................................................................................... 35
3.8 Data Visualization in R .................................................................... 35
3.9 Getting Apriori Results .................................................................... 36
3.10 Sorting Rules .................................................................................. 37
4.1 Data Ranges ..................................................................................... 42
4.2 Number of Rules ............................................................................. 44
4.3 Matching Rate HTL ........................................................................ 46
4.4 Matching Rate LTH ........................................................................ 47
4.5 Possible Types .................................................................................. 51
4.6 Occupied cells in HTL& LTH Cases .................................................. 53
4.7 Number of Sites Disagreeing Range A ............................................ 57
4.8 Number of Sites Disagreeing Range A (Continued) ....................... 58
4.9 Number of Sites Disagreeing Range A (Continued) ....................... 58
4.10 Number of Sites Disagreeing Range B .......................................... 59
4.11 Number of Sites Disagreeing Range B (Continued) ....................... 60
4.12 Number of Sites Disagreeing Range B (Continued) ....................... 60
4.13 Number of Sites Disagreeing Range C .......................................... 61
4.14 Number of Sites Disagreeing Range C (Continued) ....................... 62
4.15 Number of Sites Disagreeing Range C (Continued) ....................... 62
4.16 Outliers Results: Range A .............................................................. 68
4.17 Outliers Results: Range B .............................................................. 68
4.18 Outliers Results: Range C .............................................................. 69
4.19 Outliers Results: Range A (swap) .................................................. 69
4.20 Outliers Results: Range B (swap) ...................................... 70
4.21 Outliers Results: Range C (swap) ...................................... 70
# List of Figures

1.1 Map of Sampling Site Location [25] ......................... 5  
2.1 Data Mining Task ........................................... 8  
2.2 The Categorization of Clustering ................................ 9  
2.3 CBH Learning .................................................. 18  
2.4 Convex Hull of a Set of Points .................................. 21  
2.5 Extreme Point .................................................. 22  
2.6 Sort All the Points by Increasing Angle .......................... 22  
2.7 Graham Scan Algorithm Processing Steps ........................ 23  
2.8 Construction Dominance and Farmland Dominance .............. 25  
3.1 Apriori Results Example ......................................... 37  
3.2 Setting the Parameters .......................................... 38  
4.1 Cluster Type Calculation .......................................... 48  
4.2 Cluster Determination ........................................... 49  
4.3 Species Cluster Type ............................................ 50  
4.4 Sampling Sites Cluster Type ..................................... 50  
4.5 25%, 50% and 75% Test Matching Rate HTL & LTH, .......... 52  
4.6 Test Matching Rate with a Reduced Number of Sites .......... 54  
4.7 Stochastic Dominance: CBH & BJ result ........................ 55  
4.8 NMS Plot of sampling sites (Axes are not meaningful) [6] ... 63  
4.9 NMS Plot: CBH & BJ Result ...................................... 64  
4.10 Box Plot: CBH & BJ Result ...................................... 66  
4.11 NMS Plot: K-means & BJ Result .................................. 72  
4.12 NMS Plot: CBH & K-means ....................................... 72
Chapter 1

Introduction

1.1 Problem Statement

River ecosystems often suffer long-term degradation because of the direct or indirect impact by human beings. Pollution causes natural river quality to gradually and possibly permanently reduce [22]. The decline in river health has exacerbated the global water crisis [2]. In recent years, people have increased environmental consciousness and recognized the negative impact of changing river ecosystems. The health of a river may be influenced by many factors: water quality, geological and geomorphological composition, habitat, and biological status [2]. Protection and restoration of the river ecosystem has become a major concern. Scholars in many countries have developed a variety of methods to evaluate the health of the river. Biomonitoring, the study of macroinvertebrate populations plays an important role in river health assessment, due to the advantage of low cost and transportability of results. Aquatic macroinvertebrates are found in all waterways. Some macroinvertebrates require a high level of dissolved oxygen: the abundance of these types of macroinvertebrates may show good water quality. Otherwise, macroinvertebrates which thrive in regimes of lower dissolved oxygen often indicate polluted waters [3].
Scientists working at the China Research Academy for Environmental Sciences (CRAES) have hypothesized that there are relationships between populations of macrobenthic communities and the proportion of land surface devoted to farmland development. Based on 208 macroinvertebrate population census samples from Taizi River in north east of China, CRAES has identified the community degradation by using non-metric multidimensional scaling (NMS) [36, 25].

They have noticed that human pressures have caused significant shifts of macroinvertebrate community compositions. However, there has been little work in classification of each group based on the CRAES method, principally because there are so many sampling sites in different groups which overlap and mix with each other. It clouds the meaningfulness of the results, which renders the boundaries of the separate classifications unclear [25].

1.2 Thesis Statement

This problem had led us to see whether there are other ways to reach a determination. In order to improve the accuracy, to prepare for the introduction of new data and to provide a diagnostic framework for new assessments, we explore whether data mining technology can be used as a tool for the dataset we received.

We would expect the headwaters of a river to have the best ecosystem health and the river to become progressively degraded as one travels downstream past human activity. From this initial assumption we would find further refinement of river water quality from investigation to suggest biological or other measurement. We would also
expect improvement in river quality when human activity along the river is buffered
by riparian strips and other conservation steps.

The thesis will test the updating of the clustering of the river water quality using
measures such as biological sampling and the use of buffering. The subdivision of the
environmental quality types into three categories was imposed by CRAES analysis.
The datasets given to us are sparse and somewhat overlapping and not very large. Our
main interests are to independently reach either reinforcement of or contradiction to
CRAES original analysis. Our interests in using association rules and logical learning
stems from a desire to explore this combination of computer algorithms for clustering.

Using the apriori algorithm and current best hypothesis (CBH) search method
together is a suitable way to get more reasonable and clear clustering results in the
landuse data. We try to find all the potential relationships (rules) between samplings
and numbers of each species by using apriori algorithm and then using the current
best hypothesis algorithm training these rules in specific order to reach our goal.

Therefore, our thesis statement is to investigate the use of a data mining algo-
rithm and logical learning to cluster sparse and overlapping environmental data, and
compare the result to algorithms already in use.

1.3 Objectives

We attempt to use data mining technology classification of CRAES studies to
break the sampling sites into 3 different groups.

First, we use apriori algorithm to find so-called association rules, and based on
these rules, we organize these rules by using the CBH search algorithm, selecting
different percentage of total set of association rules as training set, and using the rest
of the rules as test set to check the training accuracy. Next, we remove irrelevant
data from the original dataset, only retaining data cells that show association rules
training results consistent with the original data, and then determine the final cluster
type in each sampling site. Finally, we compare the results of the CBH solution
with two other clustering methods, which are k-means and PC-ord software (Beijing)
results.

1.4 Research Area and Sampling Locations

The Taizi River is located in the northeast of China (122°30’ - 124°50'E and
40°30’ - 41°40'N) and belongs to the Asian temperate continental monsoon climate
zone, with a length of 413km and catchment area of 1.39 × 104 km². The main
features of the local climate are: hot rainy season (from June to August) a sunny
long cold period in winter(from November to April), and a short spring and autumn
[25].

There are 69 sampling sites in the Taizi river as of May 2009 (Figure 1.1). The
sampling sites are located both in the main channel and in subcatchments of the
river. Macroinvertebrates were collected by using a Surber net (30×30 cm, 0.5mm
mesh size) [8]. The numbers of each type of macroinvertebrate were counted under
microscopes. Based on the populations of different species, one can easily identify the
ecosystem quality level. At the subfamily level it is often the case that Chironomidae
and Limnodrilus hoffmeisteri show levels of organically polluted sampling sites [25]. The specific species names used in the CRAES study were, unfortunately, classified information which was not divulged to us. We only know them as species 1 (sp001) through species 225 (sp225).

![Figure 1.1: Map of Sampling Site Location [25]](image)

The original species table with the elimination of species whose appearance was rare (less than 1% overall) from CRAES forms a 69 stations × 99 species matrix. We can see the location of the 69 sites from Figure 1.1. Table 1.1 shows the part of the data from the Taizi River in 2009.

In this 69 × 99 matrix, there are 6831 data items in total. However, there are 5759 items with value 0 (84.31% of the total number of entries).
Table 1.1: Original Species Minus <1% Rare Species

| Sampling Sites | Sp001 | Sp003 | Sp004 | Sp005 | Sp019 | Sp020 | Sp023 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>T01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T02</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T03</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>T05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

1.5 Overview of Thesis

In Chapter 2, a literature review is presented. In this chapter, we outline and explain the background information of each component of the algorithms that we used. In Chapter 3, we demonstrate the implementation and show the main results. Then in Chapter 4, we analyze those results and compare to other known algorithms. The thesis concludes in Chapter 5 with a brief summary and directions for future research.
Chapter 2

Background and Literature Review

2.1 Data Mining and Cluster Analysis

Over the past few years, data mining research applied to many issues has received considerable research attention. For large data sets, processing and analysis is an important topic. Data mining is a multidisciplinary technology: it can extract information or patterns from a large database. Data mining tools encompass: statistics, databases, knowledge acquisition, image recognition, optimization, high-performance computing, artificial intelligence, and data graphics.

Data mining involves many complex techniques. They can be classified into two main types, predictive and descriptive, and the main features are shown in Figure 2.1 [13]. In the thesis, we focus on association rules.

There are many articles concerning data mining technology. Zaifang Zhang [35] proposed an apriori-based data mining approach and knowledge extraction from historical data. The approach is focused on mining potential useful association rules (including positive and negative rules) [35].

In the early 1980s, Michalski proposed conceptual clustering technology. Cluster analysis is a major way to divided data into meaningful groups(clusters). It can help
users to discover the relationships in large datasets. The goal of cluster analysis is to ensure the items inside of the group should have some relationship between each other, and that group items should not be related to other groups’ items. Cluster analysis plays an important role in several fields such as machine learning, image analysis, information retrieval and data mining [18].

The major approaches of clustering methods can be described as Figure 2.2. Hierarchical methods can produce a lot of partitions but partitional methods can only produce one. In the thesis, we focus on the k-means method which belongs to the partitional methods class [17]. K-means, Ward’s method [6] and support vector machine (SVM) [9] analyses were proposed and accepted in our thesis proposal as the alternatives to using our algorithms, for performance comparisons.
2.2 Discussion of Association Rules

2.2.1 Basket Analysis

A typical example of mining association rules is market basket analysis. Market analysts wish to find the relationship between different commodities that customers put into a shopping basket [5]. If a customer buys an item, how likely is that customer to also buy a related product?

This is how one extracts association rules from shopping basket data.

2.2.2 Basic Description of Association Rules

Association rules describe potential relationships among data items. The general form is “$A_1 \land A_2 \land \ldots \land A_m \Rightarrow B_1 \land B_2 \land \ldots \land B_n$” wherein $A_i (i = 1, 2, \ldots, m)$, $B_j (j = 1, 2, \ldots, n)$ are the data items and the association rules between data items is that: if items appear in a transaction, it can be deduced that other items also appear in
the same transaction [30].

**Rule Mining**

If a transaction contains X, then is this transaction likely to contain Y? The form would be: \( \{X\} \rightarrow \{Y\} \)? which can usually be described as follows. In a transaction, when the customer purchased a pen (where \( X = \text{pen} \)) then it is likely that he also purchased ink (where \( Y = \text{ink} \))? This is the association rule \( X \rightarrow Y \) [29]. This implication does not imply causality. \( Y \rightarrow X \) may also need analysis. \( X \rightarrow Y \) merely expresses the strength of the presence of Y whenever X is present. The meaning behind association rule formulation (which is the premise or premises and which is the implicand) is up to the person using the rules.

**Definition of Support, Confidence and Lift**

For the transaction set D, the rule \( X \Rightarrow Y \) is determined by support(S), confidence(C) and lift(L) constraints [20, 16].

**Support** means the frequency of occurrence in the rules. The support of an item set X is \( s(X) \), the number of transactions containing X divided by total number of transactions in D.

\[
Support(X) = \frac{\text{Number of transactions that contain } X}{\text{Total number of transactions}}
\]

**Confidence** is a measure of the strength of the rule. For the rule \( X \Rightarrow Y \) the degree of confidence is defined as: when D contains the transaction X, what is the possibility that it also contains Y. The confidence of a rule is defined as follows:
Confidence$(X \Rightarrow Y) = support(X \cup Y)/support(X)$

Lift is another measure of a rule’s quality. It is defined as

$Lift(X \Rightarrow Y) = support(X \cup Y)/support(X) \ast support(Y)$

If the lift value is greater than 1, it means that this rule is a strong rule.

The minimum support threshold indicates the minimum number of items in the item set. Minimum confidence threshold indicates a rule’s minimum reliability. If a data item set $X$ satisfies $support X \geq$ minimum support, then $X$ is a retained set of data items. This is generally given by the minimum confidence threshold and minimum support threshold. When confidence and support are greater than a corresponding threshold, and the lift value is greater than 1, a rule is called a strong association rule, otherwise it is known as a weak association rule. The task is to find association rules whose confidence and support size is greater than or equal to a specified value, and whose lift value is greater than 1 [1, 32].

2.2.3 Decomposition of Association Rules

The association rule mining problem can be decomposed into the following two sub-problems [19]:

1. Identify a transaction database $D$, and find all greater than or equal user-specified minimum support itemsets. The set of itemsets with minimum support is called the largest set of items. Itemset support means the number of items that contain these itemsets.

2. Using the maximum itemsets to generate the required association rules, for each largest itemset $A$, find all non-empty subsets $a$, if $support (A) \ \backslash \ support (a) \geq$
minconfidence, then generate association rules $a \Rightarrow (A-a)$. Support(A) \ support(a) is the confidence of rule $a \Rightarrow (A-a)$.

2.3 Apriori Algorithm

The apriori algorithm is a classic and influential algorithm for mining frequent itemsets and learning the association rules from large databases. The apriori algorithm was first proposed by Agrawal, Imielinski, and Swami [4]. The core idea is to generate a candidate set and detect those sets mining frequent itemsets [24]. This algorithm has been widely applied to commercial uses, network security and other fields [27].

In the apriori algorithm, the basic idea of finding the largest item sets is a multi-step process. First, the algorithm calculates the statistical frequency of occurrence of one item itemsets. A k-item set is an itemset that contains k items. Identify those k-itemsets whose support is not less than the desired minimum support. Then from the k-itemsets the loop continues for $k \leftarrow k + 1$ until there is no k-itemset with the desired support. The first traversal just calculates the exact value of each itemset, in order to determine the large 1 itemsets. After that, the k traversal consists of two stages. First, use $L_{k-1}$ which is found from the k-1 times traversal and candidate itemsets $C_k$ which are generated from the apriori function defined below. Then scan the database to calculate the candidates support [31].

**The pseudo code for apriori algorithm:**

Join step: $C_k$ is generated by joining $L_{k-1}$ with itself.
Prune step: Any (k-1) itemset that is not frequent cannot be a subset of a frequent k-itemset.

**Pseudo code:**

\[
C_k : \text{Candidate itemset of size } k
\]

\[
L_k : \text{frequent itemset of size } k
\]

\[
L_1 = \{\text{frequent items}\};
\]

For \(k = 1; L_k \neq \emptyset; k + +\) do begin

\[
C_{k+1} = \text{candidates generated from} L_k;
\]

For each transaction \(t\) in database do

Increment the count of all candidates in \(C_{k+1}\) that are contained in \(t\)

\[
L_{k+1} = \text{candidates in } C_{k+1} \text{ with min.support}
\]

End

Return \(U_k L_k\);

**Apriori example**

To explain the apriori concepts, we use a small example in the Taizi river database. The small data itemset contains 5 different species, sp001, sp002, sp003, sp004, sp005. The items are shown in Table 2.1. A possible rule in Taizi river could be \{sp001, sp002\} \(\Rightarrow\) \{sp003\}, which means that, if sp001 and sp002 appear, sp003 also appears.

Count the number of transactions in which each item occurs, called support (Table 2.2).
Table 2.1: Example Taizi River Database with Transactions

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sp001, sp002, sp003, sp004</td>
</tr>
<tr>
<td>2</td>
<td>sp001, sp002, sp003, sp004, sp005</td>
</tr>
<tr>
<td>3</td>
<td>sp002, sp003, sp004</td>
</tr>
<tr>
<td>4</td>
<td>sp002, sp003, sp005</td>
</tr>
<tr>
<td>5</td>
<td>sp001, sp002, sp004</td>
</tr>
<tr>
<td>6</td>
<td>sp001, sp003, sp004</td>
</tr>
<tr>
<td>7</td>
<td>sp002, sp003, sp004, sp005</td>
</tr>
<tr>
<td>8</td>
<td>sp001, sp003, sp004, sp005</td>
</tr>
<tr>
<td>9</td>
<td>sp003, sp004, sp005</td>
</tr>
<tr>
<td>10</td>
<td>sp001, sp002, sp003, sp005</td>
</tr>
</tbody>
</table>

Table 2.2: Example: Support

<table>
<thead>
<tr>
<th>Items</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp001</td>
<td>6</td>
</tr>
<tr>
<td>sp002</td>
<td>7</td>
</tr>
<tr>
<td>sp003</td>
<td>8</td>
</tr>
<tr>
<td>sp004</td>
<td>9</td>
</tr>
<tr>
<td>sp005</td>
<td>6</td>
</tr>
</tbody>
</table>

We can define the minimum support level. In Table 2.3, if we assume the minimum support = 4, then all the items are frequent. Then we create 2-pairs items.

Table 2.3: Example: Support (Continued)

<table>
<thead>
<tr>
<th>Items</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp001,sp002</td>
<td>4</td>
</tr>
<tr>
<td>sp001,sp003</td>
<td>5</td>
</tr>
<tr>
<td>sp001,sp004</td>
<td>5</td>
</tr>
<tr>
<td>sp001,sp005</td>
<td>3</td>
</tr>
<tr>
<td>sp002,sp003</td>
<td>6</td>
</tr>
<tr>
<td>sp002,sp004</td>
<td>5</td>
</tr>
<tr>
<td>sp002,sp005</td>
<td>4</td>
</tr>
<tr>
<td>sp003,sp004</td>
<td>7</td>
</tr>
<tr>
<td>sp003,sp005</td>
<td>6</td>
</tr>
<tr>
<td>sp004,sp005</td>
<td>4</td>
</tr>
</tbody>
</table>

Then we generate triple frequent items using the previous results(Table 2.4).
Table 2.4: Example: Support (Continued)

<table>
<thead>
<tr>
<th>Items</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>sp001, sp003, sp004</td>
<td>4</td>
</tr>
<tr>
<td>sp002, sp003, sp004</td>
<td>4</td>
</tr>
<tr>
<td>sp002, sp003, sp005</td>
<td>4</td>
</tr>
<tr>
<td>sp003, sp004, sp005</td>
<td>4</td>
</tr>
</tbody>
</table>

For the itemset \{sp002, sp003, sp004, sp005\}, we cannot find the threshold support, therefore, the algorithm ends.

2.4 K-means Algorithm

The K-means clustering method is another popular algorithm for clustering data sets [34]. It creates groups of individuals according to their similarity, using statistical analysis to classify them into several categories. It only works on numeric data. Also, K-means is one of the simplest learning algorithms [15]. The main idea is described as follows.

First, we define K cluster centers, where K is a specified number of clusters which must be defined in advance. The location of the initial centers could find the final results. Therefore these centers should be placed in an ingenious way. This means that a better choice is to place them as far away from each other as possible. For the rest of the points, according to their cluster center similarity (distance), allocate them respectively to its most similar cluster center and calculate the cluster center of each of the new clusters (the mean distance among all the objects in the cluster). Repeat this process until the standard measure function (eg. sum of squared distance) starts to converge [14].
Distance Measure

Hartigan and Wong [14] have defined a default distance measure algorithm in which partitions the observations into K groups and then minimizing the sum of the squares of the observed data from their designated cluster centroids, as below [14]:

$$\text{Sum}(k) = \sum_{i=0}^{n} \sum_{j=0}^{p} \left( x(i, j) - \overline{x(k, j)} \right)^2$$

$\overline{x(k, j)}$ means all j elements in group k.

Algorithm steps

Let $X = \{x_1, x_2, x_3, ..., x_n\}$ be the set of data points and $V = \{v_1, v_2, .., v_k\}$ be the set of centers.

1. Randomly select k cluster centers. The farthest point starting from an exterior point could be used for selection.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. Recalculate the new cluster centers.
5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data points were reassigned then stop, otherwise repeat from step 3.

The advantages and disadvantages of K-means algorithm [28]

Advantages:
This algorithm is fast, robust and easy to understand, and it will give good results when the clusters are distinct or separate from each other.

**Disadvantages:**

1. In the K-means algorithm, the value k is given in advance. The k value selected may be very difficult to estimate. Most times, we do not know how to divide a given data set into an appropriate number of categories.

2. In the K-means algorithm, first, we need to determine an initial partition based on the initial cluster centers, we then optimize starting from the initial partition. The choice of the initial cluster centers will affect the clustering results. If the initial value selected is bad, we may not get effective clustering results.

3. From the framework of the K-means algorithm, we can see that the algorithm needs to constantly adjust the sample classification. We continue calculating the new cluster centers after adjusting the sample classification. Therefore, when the dataset is very large, the resulting computation time might be very long.

4. Each point can only belong to a single cluster.

### 2.5 Knowledge Learning and Current Best Hypothesis Search Algorithm

A learning method can be considered as finding useful information from a hypothesis space to get a satisfactory approximation to an unknown functional relationship, and continue doing adjustment of basic assumptions in order to improve the approximate functional relationships [26].
A hypothesis space $H$ consists of all hypotheses, $H_1, \ldots, H_n$ of a set $S$. If hypothesis $H_i$ is consistent with a whole training set (a subset of $S$), then we can say it is consistent with each example in the training set. However not all hypotheses need to be consistent with an example. There are two possibilities [26].

**False negative:** In fact, a relationship may be positive, but the hypothesis says it should be negative.

**False positive:** In fact, a relationship may be negative, but the hypothesis says it should be positive.

**Current best hypothesis search algorithm**

The idea of the CBH is to adjust the hypothesis each time in order to maintain consistency with examples as they are introduced [26]. It will use generalization of the current hypothesis to include false negatives case and use specialization to exclude false positive case. If the case belongs to false negative, generalize the hypothesis by adding new disjunctions or dropping current exists terms. If the case is the part of specialization, specialize the hypothesis by dropping the disjunctions or adding new terms.

![Figure 2.3: CBH Learning](image)

---

[Image: CBH Learning]
Figure 2.3, graph (a) shows the hypothesis $H_r$ has already categorized all the positive examples (+) and negative examples (−), in graph (b), we get a new example which is negative (−) but it shows positive (+). Therefore, we should extend the hypothesis and make the new example include the hypothesis. This step is called generalization, as shown in graph (c). On the contrary, if an example is a false positive, as in graph (d), the step we should perform is to exclude this example from the hypothesis graph (e) which we call specialization.

The current best hypothesis learning algorithm shows below:

Function current-best-learning(example) returns hypothesis $H$

$H :=$ hypothesis consistent with first example

For each remaining example $e$ in examples do

If $e$ is false positive for $h$ then

$H :=$ choose a specialization of $H$ consistent with examples

Else if $e$ is false negative for $h$ then

$H :=$ choose a generalization of $H$ consistent with examples

If no consistent generalization/specialization found then fail

End

Return $H$

Specialization and generalization

The updating function can be divided into two main parts: specialization and generalization. In order to generalize, we can replace the constants with variables,
dropping conjuncts, adding disjunctions or generalizing terms.

**Specialization and generalization example:**

In the Taizi river, we count all the species that occur in the river basin. Under this circumstance, we only need to consider dropping conjuncts and adding disjunctions.

**The way of generalization:**

Dropping conjuncts:

\{sp001=1T2, sp003=3T18\} → \{sp001=1T2\}

Adding disjunctions:

\{sp001=1T2, sp003=3T18\} → \{sp001=1T2, sp003=3T18\} V \{sp023=1T2\}

**The way of specialization:**

Add conjuncts:

\{sp001=1T2, sp003=3T18\} → \{sp001=1T2, sp003=3T18\} & \{sp023=1T2\}

Dropping disjunctions:

\{sp001=1T2, sp003=3T18\} V \{sp023=1T2\} → \{sp023=1T2\}

**Difficulties and problems**

A CBH learning algorithm may have several possible ways to do the generalization or specialization steps. The extension of the hypothesis might not be the simplest hypothesis. Also, it might lead to an unrecoverable state. Under this situation, the program should backtrack [26].
2.6 Convex Hull and Graham Scan Algorithm

As an aid of visualizing the clusters of sites, we have undertaken to encircle each cluster grouping by a convex hull. The resulting picture gives us a means to compare differences in the results of the various clustering techniques.

The convex hull is the smallest convex set of points that contains all the points of a given set. The goal of a convex hull algorithm is to find the extreme vertices and link them to show the hull boundary. Figure 2.4 shows the convex hull of a set of points [10].

![Convex Hull of a Set of Points](image)

Figure 2.4: Convex Hull of a Set of Points

In order to compute the convex hull, we choose the Graham scan algorithm to reach our objective.

First, the extreme point (points [0]) should be determined, this point has the minimum y coordinate and must on the hull (Figure 2.5). Second, sort all the points by increasing the pivot angle (Figure 2.6). Then build the hull by the right hand side of the extreme point one by one and check if each point should be accepted or rejected [10].

If the point causes a left turn, then we add an edge. Otherwise, we should backtrack if the point results in a right turn. Figure 2.7 explains the steps of points
determination where the labels are P: Previous point, C: Current point and N: Next point.
Figure 2.7: Graham Scan Algorithm Processing Steps
2.7 Stochastic Dominance

For single entities (observations events or outcomes), pairwise comparison such as better, greater or larger is possible. For multiple groups of entities there will inevitably be overlaps. It becomes impossible to assign a comparative measure for distinguishing one group from another. We usually apply the principle of **stochastic dominance**.

Stochastic dominance is usually used to distinguish between decisions in areas such as analysis of income inequality or alternatives between distributions of outcomes [21]. The relationship between two distributions of entities can be measured by the cumulative distribution functions (CDF) of the measure of comparison [33]. The cumulative distribution function can be defined as:

$$P(x) = \int_{-\infty}^{x} p(t)dt$$

where $p(x)$ is the measure of the entity represented by $x$, and where $P(x)$ is the proportion of a population with value $p(.)$ less than $p(x)$.

We make the following definition. For any distribution $F$ and $G$, a CDF first-order stochastic distribution dominates (FSD) another distribution $G$ only if

$$F(x) \leq G(x) \ (\forall \ x)$$

Also, for any distribution $F$ and $G$, a CDF $F$, a second-order stochastic distribution dominates another $G$, only if

$$\int_{-\infty}^{c} F(x)dx \leq \int_{-\infty}^{c} G(x)dx$$

We use the stochastic dominance (SD) approach to analyze the factors that contribute to environmental degradation over time. The variables are: the 69 sampling sites, construction catchment and farmland catchment.
In figure 2.8 we plot the increasing percentage of farmland in each subcatchment of the Taizi river basin, for those subcatchment with “good” environmental classification, as determined by CBH HTL, and two versions of Beijing’s results. The point at which all “good” sides have been graphed occurs for our algorithm at a smaller
percentage farmland area than the result for the other two algorithms. The significance of this result will be explained in Chapter 4.
Chapter 3

Methodology and Implementation

In this chapter, we describe the algorithms we use from the perspective of the Taizi River dataset. We discuss the appropriateness of the apriori analysis for our dataset, and discuss the necessary conditions for the analysis to work. First we divide the data into different ranges based on the numbers of the species present. Then we use the apriori algorithm to create rules. We explore how different ranges and parameter settings affect the apriori results. We observe that the data standardization and sorting of training rules are necessary steps before the CBH training. After training, we test the training results and determine the final cluster type to each sampling site. Recall, the original subdivision in the training set was based on location: highlands, midlands and lowerlands. After the apriori analysis, the presence of different species will determine the actual classification of the sites in the test set.

3.1 Dataset Management

3.1.1 Scientific and Standardized Data Structure as the Basis of Apriori

All the data items in the matrix are counts of the number of particular species. The data type is integer. A zero value means there are no species of that type
occurring in the particular site. We ignore those zero values in the calculation.

Apriori has been already implemented in R [12], and it works with character data. It can automatically determine the not applicable (NA) value in a data item and ignore it. Users can specify all of the parameters that the algorithm uses to generate the association rules. We must therefore convert numeric data items into characters based on different value ranges.

### 3.1.2 Suitability of the Problem for this Study

The use of association rules for clustering depends on having sufficient data for each of the potential cluster sets to be represented. One of the central tenets of association rules is the so-called stationarity principle. Stationarity means that a subset of the data inherats all of the revelent characteristics of the whole population. In order to be able to have association rules for each cluster, we must have enough points for each potential group. There is a danger that, for a random subset of the data, some clusters will be missing. In our case there are three possible clusters, but we can develop a simple theorem for two and generalize. As the following theorem illustrates, there is a minimum number of data points from the original dataset to have some assurance of representation of both clusters. The existence of rules concerning a perticular cluster in evidence that our simple has representative items from that cluster.

We suppose there are two clusters. For clusters A and B of size r and s respectively, the probability that a single observation is in r is given by the expression

\[
Pr \ (\text{all obs in } A) = \frac{r}{r + s} = 1 - \frac{s}{r + s} = 1 - Pr(\text{obs in } B)
\]
\[ Pr (\text{two obs in A}) = 1 - Pr (\text{two obs in B}), \] but

\[ Pr (\text{two obs in B}) = Pr (\text{second obs in B} / \text{first obs in B}) \times Pr (\text{first obs in B}), \]

that is

\[ Pr (\text{two obs in B}) = \frac{(s - 1)}{(r + s - 1)} \times \frac{s}{(r + s)} \]

Continuing, for \( k \) less than \( \min s \),

\[ Pr(\text{all k obs in B}) = \frac{s(s - 1)(s - k + 1)}{(r + s - 1)(r + s - k + 1)} = \frac{s^{(k)}}{(r + s)^{k}}. \]

\[ Pr(\text{all k obs in A}) = 1 - Pr (\text{all k obs in B}) \]

We can compute that for, say 50 observations for \( r + s = n \), and for a random choice of subset, we do not “flood” the dataset by having all of the observations in one cluster subset. For certainty that at least one observation is in each set A and B(\( r + s \sim 50 \)), a quick spreadsheet calculation easily demonstrates that, for any reasonable representation of both cluster candidates, a random choice of as few as ten sample points will contain representatives of both cluster groups. In order to ensure some stability in this calculation, several experiments should be performed. If cluster B is absent from the examples, there will no “cluster B” rules generated. We generalizes this for three clusters A, B and C by applying the therom first to \( A \cup B \) and then to \( (A \cup B) \cup C \).

### 3.1.3 A Reasonable Division of the Data Range

For a priori the discrete segmentation of the range of data must be carefully performed. The number of groups can affect the a priori results. The more range groups we define, the fewer a priori results will be generated. Table 3.1 shows a
typical relationship between the number of ranges and groups generated by apriori. The amount of apriori results will affect the training step. Too many groups can result in too few rules for training. However, too few groups will cause an excessive number of rules and the results might be become rough.

Table 3.1: Ranges & Number of Rules

<table>
<thead>
<tr>
<th>ID</th>
<th>Range</th>
<th>Support/Confidence</th>
<th>Number of Rules Cluster = 1/2/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>NA, 1-2, 3-18, 19-7775</td>
<td>0.04/0.1</td>
<td>15079/255/93.</td>
</tr>
<tr>
<td>B</td>
<td>NA, 1-1, 2-5, 6-29, 30-7775</td>
<td>0.04/0.1</td>
<td>6922/195/60.</td>
</tr>
<tr>
<td>C</td>
<td>NA, 1-1, 2-4,5-14, 15-66, 67-7775</td>
<td>0.04/0.1</td>
<td>1363/85/55.</td>
</tr>
</tbody>
</table>

Recall, the number of individuals in the data matrix is 7776, and "0" occupies 84.31% of the cells. Species in the numerical range from 1 to 5 are heavily represented. According to the amount in each range and the algorithm needs, we have chosen to divide the data into 4 different ranges based on the numbers of the species present, for example:0, 1 to 2, 3 to 18, 19 to 7775. This grouping gives a nearly even distribution of cells across the data ranges.

The range can be defined as: NA, 1T2, 3T18, 19T7775, where a T b means from a individuals to b individuals.

Table 3.2: Data Range Parameter Setting

<table>
<thead>
<tr>
<th>STARTDATA</th>
<th>ENDDATA</th>
<th>CODING</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1T2</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>3T18</td>
</tr>
<tr>
<td>19</td>
<td>7775</td>
<td>19T7775</td>
</tr>
</tbody>
</table>
The user can set the parameter variables range as Table 3.2 shows. In this table, STARTDATA means “From individual number”, ENDDATA means “To individual number”, CODING means “The converted content”. For example, the first line of the table means convert all zero values into NA. The second line means convert the value 1 and 2 into characters 1T2. Table 3.3 shows an example output.

Table 3.2: Data Range

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sampling</td>
<td>cluster</td>
<td>sp001</td>
<td>sp002</td>
<td>sp003</td>
<td>sp004</td>
<td>sp005</td>
<td>sp019</td>
<td>sp020</td>
<td>sp023</td>
<td>sp029</td>
</tr>
<tr>
<td>2</td>
<td>T01</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>3</td>
<td>T02</td>
<td>1T1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>1T66</td>
</tr>
<tr>
<td>4</td>
<td>T03</td>
<td>1T4</td>
<td>NA</td>
<td>NA</td>
<td>2T4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T4</td>
<td>1T66</td>
</tr>
<tr>
<td>5</td>
<td>T04</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T4</td>
<td>1T14</td>
</tr>
<tr>
<td>6</td>
<td>T05</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T4</td>
<td>677775</td>
</tr>
<tr>
<td>7</td>
<td>T06</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>1T66</td>
<td>NA</td>
</tr>
<tr>
<td>8</td>
<td>T07</td>
<td>1NA</td>
<td>NA</td>
<td>1T1</td>
<td>1T1</td>
<td>2T4</td>
<td>1T66</td>
<td>1T1</td>
<td>1T66</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>T08</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2T4</td>
<td>NA</td>
<td>1T66</td>
<td>1T14</td>
</tr>
<tr>
<td>10</td>
<td>T09</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>NA</td>
</tr>
<tr>
<td>11</td>
<td>T10</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>2T4</td>
<td>5T14</td>
<td>NA</td>
<td>NA</td>
<td>1T4</td>
<td>2T4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>T11</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>2T4</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>T12</td>
<td>1NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>14</td>
<td>T13</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>15</td>
<td>T14</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>T15</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>T16</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>T17</td>
<td>2NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1T1</td>
</tr>
</tbody>
</table>

Table 3.3: Data Range

3.1.4 Data Standardization

Standardization of the transformed data makes it easy to read, write and count, and also to change all data into a reasonably uniform format. After the apriori calculation, we note that the data format is variable. Table 3.4 gives examples showing the apriori results output in the transformed format.

We see from Table 3.4 that the records are variable in length. We redefine the data format as below.

00001cluster=1F001{sp041=6T100}

00001cluster=2F002{sp199=101T7775 sp200=6T100}
00001cluster=3F003\{sp062=1T5 \ sp129=6T100 \ sp132=6T100 \ }

Note:

00001——Rule number. It can accommodate an array of up to 99999 positions. The row number is in fixed position: from 1 to 5 characters.

cluster——Group No. It can be cluster=2 or cluster=3 in row fixed position: from 6 to 14 characters.

F——Reserved symbols, used to determine the items that are not yet foreseeable in row fixed position: 15 characters.

001——Number of arrays in brackets, up to 999 arrays, row fixed position: from 16 to 18 characters.

\{——Data start position, row fixed position: 19 characters.

"sp041=6T100 "——Array(15 characters length). If the length of the

<table>
<thead>
<tr>
<th>Table 3.4: Format of Apriori Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 \ sp041 = 6T100 \ =&gt; \ cluster = 1 \ 0.05797101 \ 1.0000000 \ 3.0000000</td>
</tr>
<tr>
<td>1 \ {sp199 = 101T7775, \ sp200 = 6T100} \ =&gt; \ cluster = 2 \ 0.05797101 \ 1.0000000 \ 3.0000000</td>
</tr>
<tr>
<td>1 \ {sp062 = 1T5, \ sp129 = 6T100, \ sp132 = 6T100} \ =&gt; \ cluster = 3 \ 0.05797101 \ 1.0000000 \ 3.0000000</td>
</tr>
</tbody>
</table>
data item is less than 15 characters, then spaces fill the entry to 15 characters. The total length of this portion of the rule is $15 \times n$.

\[
\{ \text{——Data end position.} \}
\]

3.1.5 Items Group and Logical Operators

Table 3.4 is an example of the format. We use \{" as a start position, the end position can be determined by \}". All the data items between these two strings are identified as a group, eg. \{sp062=1T5 sp129=6T100 sp132=6T100 \} (See Table 3.4 ). The relationship among these items are logical and(\&). In other words, in this example, we can say that “sp062=1T5 \& sp129=6T100 \& sp132=6T100” is an example rule for “cluster = 3”.

For the same cluster, apriori can generate several rules, as in Figure 3.1 where the rules all point to ”cluster = 1”. The relationship of these rules is logical or(\vee).

3.1.6 Defining the Output File Name Format

We use the abbreviated name to mark output file and record the parameter setting. The name is automatically generated.

For example:

PC-ord1-s05c10p25AH.csv

PC-ord1——Final cluster output.

s05c10——Apriori parameter setting (support=0.05 confidence=0.10)

p75——Training rate = 75%
A——Range ID
H——Training method H = Training from High to Low
(Training method L = Training from Low to High)

3.2 Creation of Rules

In this section, we illustrate the use of the apriori algorithm to determine association rules with the arules package. Transactions store items labels, transaction IDs, user IDs and a binary incidence matrix. Table 3.5 shows the transactions in R syntax using landuse database [12]. The results include number of transactions(rows) and items(columns). Table 3.6 shows each item contents(items labels are generated by variable names and levels, using “=” to link) and transactions ID.

Table 3.5: Transactions in R

<table>
<thead>
<tr>
<th>R Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>landusedata.transactions &lt;- as(landusedata,&quot;transactions&quot;)</td>
</tr>
</tbody>
</table>

Table 3.6: Transaction ID and Items

<table>
<thead>
<tr>
<th>R Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>labels(landusedata.transactions)</td>
</tr>
</tbody>
</table>

The apriori algorithm employs a level-wise search for frequent itemsets. The R syntax is: apriori (data, parameter = NULL, appearance = NULL, control = NULL). The data is the object of class transactions, and it can be a binary matrix or data frame. The default parameter of the rules is: support = 0.1, confidence = 0.8.

Table 3.7 shows using transactions to list the rules. In the example we have set both the support and confidence value to 0.1.
Table 3.7: Apriori in R

```r
landusedata.rules <- apriori(Landusedata.transactions,
   parameter = list(support = 0.1, confidence = 0.1))
summary(rules)
```

We use `inspect` (object) to visualize the data (Table 3.8). We set the lift value to be greater than 1.

Table 3.8: Data Visualization in R

```r
inspect(subset(landusedata.rules, subset = lift > 1))
```

We sort the results and write them into a CSV file based on different cluster type (Table 3.9).

For our dataset, the algorithm generates three output files. Each file contains rules of a single cluster type. The apriori results can contain several rows up to thousands of rules as shown in Figure 3.1.
Table 3.9: Getting Apriori Results

```r
###Create output file: cluster = 1###
out <- capture.output(inspect(sort(subset(landusedata.rules, 
   subset = rhs %in% "cluster = 1"), by = "lift")))
write.table(out, file = file1, sep = ",", row.names = F)

###Create output file: cluster = 2###
out <- capture.output(inspect(sort(subset(landusedata.rules, 
   subset = rhs %in% "cluster = 2"), by = "lift")))
write.table(out, file = file2, sep = ",", row.names = F)

###Create output file: cluster = 3###
out <- capture.output(inspect(sort(subset(landusedata.rules, 
   subset = rhs %in% "cluster = 3"), by = "lift")))
write.table(out, file = file3, sep = ",", row.names = F)
```

### 3.3 Training and Testing Rules ###

#### 3.3.1 Sorting the Items ####

Sorting the training rules means performing an ascending sort of all the training data from fewer items to more items contained in one rule. The purpose is to simplify the process of the program and enhance the speed of data processing. Table 3.10 shows the sorting scheme.
### Before sorting:

- **00001**
  - `cluster = 1`
  - `sp148 = 1T5`
- **00002**
  - `cluster = 1`
  - `sp129 = 1T5`
  - `sp199 = 6T100`
- **00003**
  - `cluster = 1`
  - `sp096 = 6T100`

### After sorting:

- **00001**
  - `cluster = 1`
  - `sp148 = 1T5`
- **00003**
  - `cluster = 1`
  - `sp096 = 6T100`
- **00002**
  - `cluster = 1`
  - `sp129 = 1T5`
  - `sp199 = 6T100`

### 3.3.2 Effect of Sorting the Training Data

Based on the concept of the current best hypothesis search, sorting the training data by lift value is very important. The data flow is more intuitive. It can make the...
data training program simple and convenient. Also, the running time can be much shorter. Sorting the training rules does not affect the accuracy of the training results.

### 3.3.3 Training Parameter Setting

The training rate is defined in the Parameter3.csv file in cell D1. Figure 3.2 shows 25% of training. The training sort is defined in Parameter1.csv, the user can choose the sort from high to low or from low to high.

![Figure 3.2: Setting the Parameters](image)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>File1</td>
<td>File2</td>
<td>File3</td>
<td>Train%</td>
<td>Error%</td>
<td>Support</td>
<td>Confidence</td>
</tr>
<tr>
<td>1</td>
<td>101.csv</td>
<td>102.csv</td>
<td>103.csv</td>
<td>25</td>
<td>0</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 3.3.4 Testing

The main test step is to compare the training result to the test rules. The test rules are separate from the training part. For example, if the user select a 25% training rate, then the remaining 75% should be used for testing. If the coincidence rate between training results and test rules is high enough, then we can identify the training result as “good”. The coincidence means when the user compares all the training results to one test rule, if the test rules cluster type is the same as the cluster type calculated from the training results, we conclude that, for this specific test rule,
the training result prediction is accurate.

**Training results data and test data comparison**

If the training result and test rule are identical, or the training result contains the test rule, then the test rule belongs to the range of training results. In this case, we can conclude that the results match. Otherwise, the rules result does NOT match.

Examples shows below:

**Match:**  
Training results: A  
Test Rule: A ∧ B

**Not Match:**  
Training results: A ∧ B  
Test Rule: A ∧ C

**Matching example:**

Training result: 00010cluster=3F001\{sp148=1T5 \}  
Test rule: 00059cluster=3F002\{sp148=1T5 sp129=6T100 \}

**Not matching example:**

Training result: 00055cluster=3F002\{sp081=6T100 sp129=6T100 \}  
Test rule: 00066cluster=3F002\{sp081=6T100 sp128=101T7775 \}

### 3.4 Determining the Sampling Sites Cluster Group

#### 3.4.1 Data Selection and Cluster Determination

Data selection keeps the training results data corresponding to the location of the source data table, and empties the content of the rest of the data cells. In other words, we only keep the data cells which align with the training results and mark the type of training results. For example, we mark “A” into all the cells which are related
to cluster = 1. We mark all cells related to cluster = 2 as “B” and mark the training results related to cluster = 3 cells as “C”. Then we count the total number of A, B and C in each sampling sites. The largest number we define as the final cluster type.

This environmental example (river site clustering) has significant overlap in the outcome. We may have species which indicate the site is of cluster type A, and species which say it is cluster B. we could possibly leave the results ambiguous (both A and B). For this application, we have further refined the clustering to overcome this ambiguity, placing the site in a unique group based on the assumptions that particular evidence might favor eg. a “worst case” outcome.

For some specific cases, we determine the cluster type as follows:

If $A = B > C$, we consider the final cluster as B, which is cluster = 2
If $B = C > A$, we consider the final cluster as C, which is cluster = 3
If $A = C > B$, we consider the final cluster as B, which is cluster = 2
If $A = B = C$, we consider the final cluster as B, which is cluster = 2

3.5 Summary of Chapter 3

The extension of the analysis is intended for produce a unique result, rather than having some sites belong to more than one cluster. Note that most clustering algorithms, unlike this one, place every item in a single cluster.

Since we wish to err on the pessimistic side, we have decided that, if “cluster = 3” is being asserted, it is the dominant result. The rule $A = B = C$ is one example for which the evidence appears contradictory, and therefore “cluster = 2” is chosen.
Chapter 4

Evaluation

We can now recap the previous two chapters’ considerations. There are two main transformations of the original $69 \times 99$ matrix. One is the choice of range settings, which decides the individual number’s of each species and the other is the apriori algorithm parameter setting, such as support, confidence and lift value.

The numeric data is converted into characters based on individual range of each species. These ranges are used as source dataset, and apriori algorithm is applied to find the association rules. We have three different group ranges. Each group’s apriori results are collected. From those results we find the relationship between range setting and the number of association rules.

We have that different apriori parameter settings can affect the results enormously. For each group of ranges, we have tested a range of support and confidence values to determine the relationship between the derived association rules result and the apriori parameter settings.

The training results with good test matching rate (generally good results) are selected. We compute the sampling sites’ final clusters based on their individual species classification and finally, we summarize the statistics.

The Graham scan algorithm is applied in our program to visually display the
overlaps of different cluster sampling sites. We list each range of group’s overlaps
sampling sites and illustrate the difference.

Finally, we examine the case for cluster using stochastic dominance.

4.1 Landuse Data Statistics

Landuse data [25] sampling sites can be classified into three main groups based
on natural watershed boundaries: mountain forest areas (highlands), hilly forest ar-
eas (midlands), and agricultural areas (Lowlands).

The highlands include sampling sites:

T01, T02, T03, T04, T05, T06, T07, T08, T10, T11, T12, T25, T26, T27,
T29, T65, T66, T67, T68, T69, T70.

The midlands contain:

T09, T13, T14, T15, T16, T17, T18, T19, T20, T21, T22, T23, T24, T28,
T34, T35, T36, T37, T38, T62.

The Lowlands include:

T31, T32, T33, T39, T40, T41, T42, T43, T44, T45, T46, T47, T48, T49,
T50, T51, T52, T53, T54, T55, T56, T57, T58, T59, T60, T61, T63.

<table>
<thead>
<tr>
<th>Range ID</th>
<th>Number of groups(exclude 0)</th>
<th>Group Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>NA, 1-2, 3-18, 19-7775</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>NA, 1-1, 2-5, 6-29, 30-7775</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>NA, 1-1, 2-4, 5-14, 15-66, 67-7775</td>
</tr>
</tbody>
</table>

Here are 6831 cells with 5759 zero, 240 with single individual. There are 625 cells
with fewer than 10 individuals. Between 10 to 100 we have 299 cells; from 101 to 999 individuals are present in 123 and only 25 cells contain more than 1000 individuals. This is a very sparse matrix. We use individual mean values to determine groups. The number of groups and the corresponding range segments are seen the table 4.1.

4.2 Analysis of Apriori Results

The group selection affects the apriori results (Table 4.2). If we hold the same support and confidence value, and increase the number of groups, the numbers of transactions and the rules decrease. In this analysis, we include 3 different groups of range (excluding zero value). One or two group ranges could not be selected because information is too general. Although six and more groups’ information trends to more detail, more groups could not generate enough rules through apriori method, making it hard to find suitable association rules. Recall that three variables which affect the apriori results: range selection, support and confidence value.

As we check three different range results, we can conclude the confidence value does not affect the number of rules. However, the support value affects the results significantly. When we lower the support value, more rules can be generated. In our case, we ignore support values lower than 0.04, because the similarity is lower than our expectations and information is too general. Although the support value greater than 0.06 can generate high similarity results. Anything fewer than 100 rules are too few and it is hard to get accurate predictions. We use support =0.4, 0.5, 0.6 as an example, showing the results as Table 4.2.
Table 4.2: Number of Rules

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>S=0.04, C=0.1</td>
<td>189515</td>
<td>15079</td>
<td>255</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25905</td>
<td>2389</td>
<td>92</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.05</td>
<td>25905</td>
<td>2389</td>
<td>92</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>8248</td>
<td>829</td>
<td>36</td>
<td>21</td>
</tr>
<tr>
<td>B</td>
<td>S=0.04, C=0.1</td>
<td>83620</td>
<td>6922</td>
<td>195</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>7271</td>
<td>737</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.05</td>
<td>7271</td>
<td>737</td>
<td>53</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>2481</td>
<td>275</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>C</td>
<td>S=0.04, C=0.1</td>
<td>12848</td>
<td>1363</td>
<td>85</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>2301</td>
<td>261</td>
<td>37</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.05</td>
<td>2301</td>
<td>261</td>
<td>37</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>816</td>
<td>97</td>
<td>20</td>
<td>16</td>
</tr>
</tbody>
</table>

Overall, as we can see from Table 4.2, we have a large number of rules in cluster = 1, cluster = 2 tends to a medium number and the minimum number of rules belongs to cluster = 3.
4.3 Current Best Hypothesis Analysis

4.3.1 Sorting Method

We sort the association rules using apriori lift value [7]. We train the rules HTL and LTH. We discard the rules when the lift value drops down to below 1. Training different range lift values results in totally different rules, likely due to the sparsity of the data. However, when training using 100% of rules, the results are the maintained the same in either HTL or LTH. This is not surprising, because our CBH algorithm should work the same in both directions when all the association rules are present. To see the difference between training HTL and LTH lift value, please see the table 4.3 and table 4.4.
<table>
<thead>
<tr>
<th>ID</th>
<th>Support=, Confidence=</th>
<th>Training% By HTL</th>
<th>Training rules cluster =1/2/3</th>
<th>Test rules cluster =1/2/3</th>
<th>Test matching rules number</th>
<th>Test matching rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>3770/64/23</td>
<td>11309/191/70</td>
<td>10178/11/3</td>
<td>89.99/5.759/4.285</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>7540/128/46</td>
<td>7539/127/47</td>
<td>6408/12/18</td>
<td>84.99/9.448/38.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>11309/191/70</td>
<td>3770/64/23</td>
<td>2639/10/7.</td>
<td>70/15.62/30.43</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>597/23/10</td>
<td>1792/69/32</td>
<td>1306/2/0</td>
<td>72.87/2.898/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>1194/46/21</td>
<td>1195/46/21</td>
<td>709/3/6</td>
<td>59.33/6.521/28.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>1792/69/32</td>
<td>597/23/10</td>
<td>111/1/6</td>
<td>18.59/4.347/60</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>207/9/5</td>
<td>622/27/16</td>
<td>379/0/0</td>
<td>60.93/0/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>414/18/10</td>
<td>415/18/11</td>
<td>172/0/3</td>
<td>41.44/0/27.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>622/27/16</td>
<td>207/9/5</td>
<td>42/0/3</td>
<td>28.20/0/60</td>
</tr>
<tr>
<td>B</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>1730/49/15</td>
<td>5192/146/45</td>
<td>4696/8/0</td>
<td>90.44/5.479/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>3461/98/30</td>
<td>3461/97/30</td>
<td>2965./8/12</td>
<td>85.66/8.247/40</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>5192/146/45</td>
<td>1730/49/15</td>
<td>1234/3/7</td>
<td>71.32/6.122/46.66</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>184/13/7</td>
<td>553/40/22</td>
<td>308/1/2</td>
<td>55.69/2.5/9.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>368/26/14</td>
<td>369/27/15</td>
<td>142/3/3</td>
<td>38.48/11.11/20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>553/40/22</td>
<td>184/13/7</td>
<td>27/0/1</td>
<td>14.67/0/14.28</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>69/8/4.</td>
<td>206/22/10</td>
<td>63/0/2</td>
<td>30.58/0/20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>138/15/7</td>
<td>137/15/7</td>
<td>8/0/1</td>
<td>5.839/0/14.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>206/22/10</td>
<td>69/8/4.</td>
<td>11/1/0</td>
<td>15.94/12.5/0</td>
</tr>
<tr>
<td>C</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>341/21/14</td>
<td>1022/64/41</td>
<td>693/2/0</td>
<td>67.8/3.125/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>682/42/28</td>
<td>681/43/27</td>
<td>411/2/1</td>
<td>60.35/4.651/3.703</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>1022/64/41</td>
<td>341/21/14</td>
<td>71/2/6.</td>
<td>20.82/9.523/42.85</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>65/9/6.</td>
<td>196/28/20</td>
<td>57/2/0</td>
<td>29.08/7.142/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>130/18/13</td>
<td>131/19/13</td>
<td>18/0/0</td>
<td>13.74/0/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>196/28/20</td>
<td>65/9/6.</td>
<td>4/1/2.</td>
<td>6.153/11.11/33.33</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>24/5/4.</td>
<td>73/15/12</td>
<td>5/1/0</td>
<td>6.849/6.666/0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>48/10/8.</td>
<td>49/10/8.</td>
<td>3/1/1.</td>
<td>6.122/10/12.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>73/15/12</td>
<td>24/5/4.</td>
<td>1/0/2</td>
<td>4.166/0/50</td>
</tr>
<tr>
<td>ID</td>
<td>Support=</td>
<td>Confidence=</td>
<td>Training%</td>
<td>Training rules</td>
<td>Test rules</td>
<td>Test matching</td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>-------------</td>
<td>-----------</td>
<td>----------------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>By LTH</td>
<td>cluster =1/2/3</td>
<td>cluster =1/2/3</td>
<td>rules number</td>
</tr>
<tr>
<td>A</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>3770/64/23</td>
<td>11309/191/70</td>
<td>11273/155/33</td>
<td>99.68/81.15/47.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>7540/128/46</td>
<td>7539/127/47</td>
<td>7503/115/25</td>
<td>99.52/90.55/53.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>11309/191/70</td>
<td>3770/64/23</td>
<td>3734/59/20</td>
<td>99.04/92.18/86.95</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>597/23/10</td>
<td>1792/69/32</td>
<td>1786/51/4</td>
<td>99.66/73.91/12.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>1194/46/21</td>
<td>1195/46/21</td>
<td>1189/34/11</td>
<td>99.49/73.91/52.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>1792/69/32</td>
<td>597/23/10</td>
<td>591/18/8</td>
<td>98.99/78.26/80</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>207/9/5</td>
<td>622/27/16</td>
<td>410/11/5</td>
<td>98.79/61.11/45.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>414/18/10</td>
<td>415/18/11</td>
<td>410/11/5</td>
<td>98.91/62.96/52.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>622/27/16</td>
<td>207/9/5</td>
<td>202/5/4</td>
<td>97.58/55.55/80</td>
</tr>
<tr>
<td>B</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>1730/49/15</td>
<td>5192/146/45</td>
<td>5156/120/15</td>
<td>92.3/82.19/33.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>3461/98/30</td>
<td>3461/97/30</td>
<td>3425/85/8</td>
<td>98.95/87.62/26.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>5192/146/45</td>
<td>1730/49/15</td>
<td>1694/41/10</td>
<td>97.91/83.67/66.66</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>184/13/7</td>
<td>553/40/22</td>
<td>540/26/8</td>
<td>97.64/65.36/36.36</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>368/26/14</td>
<td>369/27/15</td>
<td>365/17/4</td>
<td>98.91/62.96/26.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>553/40/22</td>
<td>184/13/7</td>
<td>180/9/3</td>
<td>97.82/69.23/42.85</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>69/8/4.</td>
<td>206/22/10</td>
<td>180/15/3</td>
<td>87.37/68.19/30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>138/15/7</td>
<td>137/15/7</td>
<td>134/10/1</td>
<td>97.81/66.66/14.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>206/22/10</td>
<td>69/8/4.</td>
<td>66/5/2.</td>
<td>95.65/62.5/50</td>
</tr>
<tr>
<td>C</td>
<td>S=0.04, C=0.1</td>
<td>25%</td>
<td>341/21/14</td>
<td>1022/64/41</td>
<td>990/34/15</td>
<td>96.86/53.12/36.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>682/42/28</td>
<td>681/43/27</td>
<td>649/29/18</td>
<td>95.3/67.44/66.66</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>1022/64/41</td>
<td>341/21/14</td>
<td>309/16/10</td>
<td>90.61/76.19/71.42</td>
</tr>
<tr>
<td></td>
<td>S=0.05, C=0.1</td>
<td>25%</td>
<td>65/9/6.</td>
<td>196/28/20</td>
<td>154/15/5</td>
<td>78.57/53.57/25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>130/18/13</td>
<td>131/19/13</td>
<td>122/13/9</td>
<td>93.12/68.42/69.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>196/28/20</td>
<td>65/9/6.</td>
<td>56/6/4.</td>
<td>86.15/66.66/66.66</td>
</tr>
<tr>
<td></td>
<td>S=0.06, C=0.1</td>
<td>25%</td>
<td>24/5/4.</td>
<td>73/15/12</td>
<td>61/8/3.</td>
<td>83.56/53.33/25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>48/10/8.</td>
<td>49/10/8.</td>
<td>43/6/4.</td>
<td>87.75/60/50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>73/15/12</td>
<td>24/5/4.</td>
<td>19/3/3.</td>
<td>79.16/60/75</td>
</tr>
</tbody>
</table>
4.3.2 Sampling Sites Cluster Type

To get the sites clustered, we add all individual species clusters in the same site and calculate the site type. Here are the factors which affect the results: group range, support value, training rate, training by lift, and choice of HTL or LTH. We have used Range A Support =0.05 Confidence =0.1, sort HTL,75% training case as an example (Figure 4.1). The final cluster type is determined by the maximum number of species of the specific class. The X-axis is the sampling sites number. The Y-axis label is the total number of individual species in the site, by cluster. The final cluster type for each cell is shown by the black line in the figure 4.1.

![Figure 4.1: Cluster Type Calculation](image-url)
Figure 4.2: Cluster Determination

**Multiple type of species**

For the species cluster results, some of the data cells may be assigned to more than one cluster. For example, figure 4.3 cell AE18 shows cluster type AB, meaning both cluster = 1 and cluster = 2. Table 4.5 enumerates the 7 different possible types in total.

**Multiple types of sampling sites**

When the total population is the same for 2 or more sampling stations (as the case with T39, T48 and T49 in Figure 4.4, the cluster type is going to be the same for all (cluster 3 in this case).

In the BJ analysis, we have also sites which correspond to the least disturbed area which include the highest taxa and individual population counts.
If a sampling site can be in either of two clusters, we automatically assign it to the more polluted one using the rules defined in section 3.4.
Table 4.5: Possible Types

<table>
<thead>
<tr>
<th>Possible types</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Cluster = 1</td>
</tr>
<tr>
<td>B</td>
<td>Cluster = 2</td>
</tr>
<tr>
<td>C</td>
<td>Cluster = 3</td>
</tr>
<tr>
<td>AB</td>
<td>Cluster = 1 or 2</td>
</tr>
<tr>
<td>AC</td>
<td>Cluster = 1 or 3</td>
</tr>
<tr>
<td>ABC</td>
<td>Cluster = 1 or 2 or 3</td>
</tr>
</tbody>
</table>

4.3.3 Training Results & Test Matching Rate

In high to low lift value cases:

Training the rules in high lift value means training the rules with high similarity first. When the training rate increases, the scope of the training results will increase (Please see appendix B). However, the test matching rate is low because the remaining part of the lift value is very low (low similarity in the rules). Figure 4.5 shows range A, HTL case matching rate. 101, 102, 103 are the files name, which means all rules related to cluster = 1 is under file 101, and so on. 25%, 50%, 75% means training rate.

In low to high lift value cases:

For the low lift value training, we train low similarity rules first. As with high to low lift value case, when we increase the training rate, the range of the training results becomes greater (please see appendix B). The test matching rate in this case is quite high because the test rules we use have high lift value (high similarity). Figure 4.5 shows range A, LTH case matching rate.

Number of cells occupied:

There is another way we can conclude how widely the range of training affects
Figure 4.5: 25%, 50% and 75% Test Matching Rate HTL & LTH,

the results. Table 4.6 shows how many cells have been selected by using training results (cells which match the training results). We generate two cases, from high to low (HTL) and from low to high (LTH). In Range A, keeping the same support and confidence value and training rate, be find fewer nonzero cells when we use the HTL case. When we increase the support value or training rate, the results are the same. Range A and Range C situations are similar. We can conclude that if the range, support, confidence and training rate are all the same, the CBH training of rules by using HTL lift value, the number of cells occupied is always fewer than with training rules by using LTH lift case.
Table 4.6: Occupied cells in HTL & LTH Cases

<table>
<thead>
<tr>
<th>Range ID</th>
<th>Support=, Confidence=</th>
<th>Training%</th>
<th>Cells occupy HTL Cluster =1/2/3</th>
<th>Cells occupied LTH Cluster =1/2/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04, 0.1</td>
<td>25%</td>
<td>374/103/66</td>
<td>629/318/196</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>374/215/106</td>
<td>629/436/275</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>374/311/185</td>
<td>629/467/370</td>
</tr>
<tr>
<td></td>
<td>0.05, 0.1</td>
<td>25%</td>
<td>237/92/52</td>
<td>507/264/144</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>237/185/112</td>
<td>507/333/208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>237/290/202</td>
<td>507/393/289</td>
</tr>
<tr>
<td></td>
<td>0.06, 0.1</td>
<td>25%</td>
<td>178/59/47</td>
<td>395/176/96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50%</td>
<td>178/99/72</td>
<td>431/233/150</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75%</td>
<td>188/199/170</td>
<td>431/280/214</td>
</tr>
</tbody>
</table>

4.4 Reduction in Numbers of Sampling Sites for Training

Compared with all 69 sampling sites CBH analysis, we randomly select 50% of total sampling sites by each cluster group, and using the same methods, run the program to check the test matching rate (Figure 4.6). Here is the LTH Range A, support =0.05, confidence =0.1.

In LTH, 50% of original sampling sites matching rate in Cluster =1 are similar to the 100% matching results. The cluster = 2 and 3 results using 50% of randomly selected sampling sites’ gives a matching rate higher than for 100% of sites.

4.5 Dominance Analysis

Figure 4.7 shows the farmland (catchment) dominance analysis in different range selection, S=0.05 C=0.1, 75% of training in Cluster =1 (section 2.7). The blue lines (Beijing 1) mean the original Beijing results, the red lines (Beijing2) are the results when we swap cluster 1 and 2 from Beijing1. The black lines are CBH HTL or CBH
Figure 4.6: Test Matching Rate with a Reduced Number of Sites

LTH. These are the results using apriori rules with CBH training.

We see that, for lower percentage of agriculture land use, the original Beijing results underpredict the number of “good” sites, are optimistic in the higher (24%) percentage. Our results an the Beijing 2 results agree very closely, with 100% of the category of sites, reached at around 24%. This result strengthens our conjecture that the original Ward’s method in PC-ord is fundamentally wrong, as opposed to our results.
Figure 4.7: Stochastic Dominance: CBH & BJ result
4.6  Current Best Hypothesis vs. Beijing Results

4.6.1  Number of Sites Disagreeing

Here are the differences between Beijing results (original) and CBH (HTL/ LTH) cases. Table 4.7 to Table 4.15 shows the number of sampling sites occurring in Beijing results and CBH results. For example, we get 19 sampling sites in Beijing original results showing cluster = 1 and in the CBH method HTL, analysis shows cluster = 2 to be similar to Beijing’s cluster = 1.

We use training rate 75%, support = 0.04 confidence = 0.1, Range A. As a comparison, we swapped Beijing cluster = 1 and cluster = 2 value, and maintained the rest of range and parameters same (Table 4.7).

We next use training rate 75%, Support = 0.05 Confidence = 0.1, Range A and with swapped Beijing results (Table: 4.8).

Then we use training rate 75%, Support = 0.06 Confidence = 0.1, Range A and with swapped Beijing results (Table: 4.9).

By calculating the CBH results compared with original Beijing results, we find that when we train with the LTH case, we have a few sampling sites showing a big difference, where Beijing results says “good” but CBH results indicate “bad”. However, if we swap the Beijing cluster = 1 and cluster = 2, the Beijing good and CBH bad cases drop down to zero in both HTL and LTH.

When we increase the support value from 0.04 to 0.06, the numbers of Beijing Good and CBH bad sampling sites are increased. However, if we swap the Beijing cluster 1 and 2, the numbers of Beijing Good and CBH bad sampling sites still match
well in both HTL and LTH cases.

The important information in the tables is seen in the diagonal entries in the bottom half of each table, in boldface. These values show agreement between good, middle and bad for Beijing swap and CBH (HTL and LTH).

Table 4.7: Number of Sites Disagreeing Range A

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>4</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>19</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>10</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>14</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>14</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>5</td>
<td>0</td>
<td>11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>15</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>4</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>10</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>18</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>1</td>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>
Table 4.8: Number of Sites Disagreeing Range A (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>1</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>17</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>15</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>14</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>17</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>2</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>15</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>4</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>18</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>1</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>2</td>
<td>8</td>
</tr>
</tbody>
</table>

Table 4.9: Number of Sites Disagreeing Range A (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>1</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>15</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>17</td>
<td>0</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>16</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>13</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>4</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>5</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>17</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>19</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>0</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>
We also test range B (The intermediate or cluster = 2 sites). There are differences between Beijing results (Original/Swap) and CBH (HTL/LTH) cases.

We use training rate 75%, Support = 0.04 Confidence = 0.1, Range B. and swap Beijing results (Table: 4.10).

Next, we use training rate 75%, Support = 0.05 Confidence = 0.1, Range B. and swap Beijing results (Table: 4.11).

Then we use training rate 75%, Support = 0.06 Confidence = 0.1, Range B. and swap Beijing results (Table: 4.12).

Table 4.10: Number of Sites Disagreeing Range B

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>3</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>15</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>15</td>
<td>0</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>12</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>15</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>6</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td><strong>15</strong></td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>4</td>
<td><strong>15</strong></td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>15</td>
<td><strong>17</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td><strong>18</strong></td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>1</td>
<td><strong>15</strong></td>
<td>5</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>6</td>
<td><strong>10</strong></td>
</tr>
</tbody>
</table>
### Table 4.11: Number of Sites Disagreeing Range B (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>2</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>22</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>9</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>19</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>12</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>2</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>4</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>9</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>18</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>1</td>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

### Table 4.12: Number of Sites Disagreeing Range B (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>1</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>27</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>4</td>
<td>0</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>16</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>14</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>5</td>
<td>27</td>
<td>2</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>4</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>19</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>0</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
We also test range C (The bad or cluster = 3 sites). There are differences between Beijing results (Original/Swap) and CBH (HTL/LTH) cases.

We use training rate 75%, Support = 0.04 Confidence = 0.1, Range C. and swap Beijing results(Table: 4.13)

Next, we use training rate 75%, Support = 0.05 Confidence = 0.1, Range C. and swap Beijing results(Table: 4.14)

Then we use training rate 75%, Support = 0.06 Confidence = 0.1, Range C. and swap Beijing results(Table: 4.15)

<table>
<thead>
<tr>
<th>Table 4.13: Number of Sites Disagreeing Range C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HTL results/Beijing_O</strong></td>
</tr>
<tr>
<td>Good (cluster=1)</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
</tr>
</tbody>
</table>

| **LTH results/Beijing_O** | Good (cluster=1) | Mid (cluster=2) | Bad (cluster=3) |
| Good (cluster=1) | 14 | 18 | 1 |
| Mid (cluster=2) | 15 | 1 | 6 |
| Bad (cluster=3) | 4 | 0 | 10 |

| **HTL results/Beijing_S** | Good (cluster=1) | Mid (cluster=2) | Bad (cluster=3) |
| Good (cluster=1) | 15 | 3 | 0 |
| Mid (cluster=2) | 3 | 19 | 1 |
| Bad (cluster=3) | 1 | 11 | 16 |

| **LTH results/Beijing_S** | Good (cluster=1) | Mid (cluster=2) | Bad (cluster=3) |
| Good (cluster=1) | 18 | 14 | 1 |
| Mid (cluster=2) | 1 | 15 | 6 |
| Bad (cluster=3) | 0 | 4 | 10 |
Table 4.14: Number of Sites Disagreeing Range C (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>2</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>15</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>16</td>
<td>1</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>10</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>20</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>3</td>
<td>0</td>
<td>9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>3</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>1</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>18</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>1</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

Table 4.15: Number of Sites Disagreeing Range C (Continued)

<table>
<thead>
<tr>
<th>HTL results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>3</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>14</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>15</td>
<td>0</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_O</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>7</td>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>19</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>6</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HTL results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>16</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>3</td>
<td>14</td>
<td>0</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LTH results/Beijing_S</th>
<th>Good (cluster=1)</th>
<th>Mid (cluster=2)</th>
<th>Bad (cluster=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good (cluster=1)</td>
<td>19</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Mid (cluster=2)</td>
<td>0</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Bad (cluster=3)</td>
<td>0</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>
4.7 Two Dimensional Plot

The plot of the high-dimensional matrix is simple; it was already implemented in PC-ord software. It can reduce high dimensional visualization into a 2 dimensional plot [6]. It shows the 69 samplings sites as Figure 4.9.

We use PC-ord 6.0 software to plot the sampling sites by using non metric multi-dimensional scaling (NMS) [6]. We use Sorensen (Bray-Curtis) Distance Measure, not penalized (Kruskal’s primary approach) [11]. Parameter setup: Number of axes \( k = 3 \), Number of runs with real data = 5, stability criterion = 0.00001, iterations to evaluate stability = 15 Maximum number of iterations = 250, step down in dimensionality, Initial step length = .20, Random numbers to starting coordinates and varimax rotation. User supplied seed integer = 1. Figure 4.8 shows the NMS plot. This layout is a two dimensional projection of the multi-dimensional cluster space.

Figure 4.8: NMS Plot of sampling sites (Axes are not meaningful) [6]
In Current best Hypothesis, we graph the CBH HTL case, by adding the Graham scan convex hull method [10], it is easy to see that three groups are almost separated from each other, with only a few points mixed, and that the visualizations of the two results are very similar.

![Figure 4.9: NMS Plot: CBH & BJ Result](image)

### 4.7.1 Box Plot

We use a box plot to show the landuse data set’s lowest, highest and median values. The first row of box plots is drawn using results of the CBH algorithm (Range A training by LTH support = 0.04 confidence = 0.1 training 75 percentage). The second row box plots shows Beijing original results. Overall, the biggest difference between these two methods is the median value. In our method, the median value of cluster = 2 is greater than cluster = 1’s in all 4 cases(construction % catchment, Farmland% catchment, construction % buffer, and Farmland % buffer). However, in
Beijing's results, the median value of cluster = 2 is lower than for cluster = 1. In both methods the cluster = 3 median value is higher than that for cluster = 1 and cluster = 2. If we look at the sampling sites on the top of the graphic map Figure 1.1, we see that the sites T2, T3, T4, T5, T6, T7, T8 are located in high mountain area (forest area). The construction and farmland (catchment or buffer) should be lower than for hilly and plains area. But in [25], they assign T2 to T8 (and more) sites to cluster = 2. Coincidentally, T26 to T35 are located in the midlands which is in a hilly area (near to urban lowlands on the map). They assign those sites to cluster = 1.

Although the numbers of macroinvertebrates are not sufficient to explain the water quality, they give an important indicator. Regional differences in the water quality are always a controversial topic. However, in our method, under Range A, S = 0.05, C = 0.1, Training 75% by lift value HTL case, T2 to T8 all show cluster = 1, T26 to T35 show: 1, 1, 2, 2, 3, 3, 3, 3, 3, 2, respectively. These results compared to Beijing's results show increased consistency with geographical trends.

However in the LTH case, Figure 4.10, we get fewer outliers in each group. Comparing HTL, the median value has slightly increased, but the trends are not changed.
Figure 4.10: Box Plot: CBH & BJ Result
4.7.2 Outliers Analysis

In order to detect the outliers between Beijing cluster results and CBH method, we combine cluster = 1 and cluster = 2 together as a light pollution group, and leave cluster = 3 as the heavily pollution group. We get outliers as follows:

Case 1: Original Beijing results VS CBH results (Table 4.16, Table 4.17, Table: 4.18).

Case 2: Swapped original Beijing cluster =1 and cluster =2. V.S. CBH results (Table 4.19, Table 4.20, Table: 4.21).

In HTL training method, as the training rate is increased, both case 1 and case 2 outliers are gradually increasing in numbers in all three ranges. On the contrary, the outliers in both case 1 and case 2 are decreased in LTH case, when we increase the training percentage.
Table 4.16: Outliers Results: Range A

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training %</th>
<th>Outlier %</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>15.942029</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>15.2173913</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>17.3913043</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>21.7391304</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>10.1449275</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>18.8405797</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>25.3623188</td>
</tr>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>15.2173913</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>12.3188406</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>10.1449275</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>15.942029</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>13.0434783</td>
</tr>
</tbody>
</table>

Table 4.17: Outliers Results: Range B

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training %</th>
<th>Outlier %</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>15.942029</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>22.4637681</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>21.7391304</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>17.3913043</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>19.5652174</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>25.3623188</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>20.2898551</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>07.2463768</td>
</tr>
<tr>
<td>B</td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>13.7681159</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>15.2173913</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>16.6666667</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>12.3188406</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>15.2173913</td>
</tr>
</tbody>
</table>
Table 4.18: Outliers Results: Range C

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training%</th>
<th>Outlier%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>15.2173913</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>18.115942</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>17.3913043</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>16.6666667</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>23.9130435</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>16.6666667</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>21.7391304</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>16.6666667</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>12.3188406</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>11.5942029</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>13.7681159</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>07.9710145</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>09.4202899</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>21.0144928</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>16.6666667</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>14.4927536</td>
</tr>
</tbody>
</table>

Table 4.19: Outliers Results: Range A (swap)

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training%</th>
<th>Outlier%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>06.5217391</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>07.9710145</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>07.9710145</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>09.4202899</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>08.6956522</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>10.8695652</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>07.2463768</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>12.3188406</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>13.0434783</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>15.2173913</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>13.7681159</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>08.6956522</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>14.4927536</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>12.3188406</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>08.6956522</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>15.942029</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>13.7681159</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>10.1449275</td>
</tr>
</tbody>
</table>
### Table 4.20: Outliers Results: Range B (swap)

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training%</th>
<th>Outlier%</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>07.9710145</td>
</tr>
<tr>
<td>B</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>12.3188406</td>
</tr>
<tr>
<td>B</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>10.8695652</td>
</tr>
<tr>
<td>B</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>12.3188406</td>
</tr>
<tr>
<td>B</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>12.3188406</td>
</tr>
<tr>
<td>B</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>06.5217391</td>
</tr>
<tr>
<td>B</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>15.942029</td>
</tr>
<tr>
<td>B</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>13.7681159</td>
</tr>
<tr>
<td>B</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>04.3478261</td>
</tr>
</tbody>
</table>

### Table 4.21: Outliers Results: Range C (swap)

<table>
<thead>
<tr>
<th>Range</th>
<th>Support</th>
<th>Confidence</th>
<th>HTL/LTH</th>
<th>Training%</th>
<th>Outlier%</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>08.6956522</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>10.1449275</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>10.1449275</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>08.6956522</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>09.4202899</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>13.0434783</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>25</td>
<td>09.4202899</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>50</td>
<td>07.2463768</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>HTL</td>
<td>75</td>
<td>10.8695652</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>16.6666667</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>11.5942029</td>
</tr>
<tr>
<td>C</td>
<td>0.04</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>08.6956522</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>13.7681159</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>06.5217391</td>
</tr>
<tr>
<td>C</td>
<td>0.05</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>07.2463768</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>25</td>
<td>18.8405797</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>50</td>
<td>15.942029</td>
</tr>
<tr>
<td>C</td>
<td>0.06</td>
<td>0.1</td>
<td>LTH</td>
<td>75</td>
<td>10.1449275</td>
</tr>
</tbody>
</table>
4.8 Current Best Hypothesis vs. K-means

As outlined in our thesis proposal, we apply the k-means method by using the arules [12] package as implemented in R. We use nstart, a parameter that lets k-means algorithm run many random times and kept the best results. We used nstart = 20 as example. Three groups have a large overlap and fewer cluster = 1 members are generated. However, we can easily see that the sites located in the highlands were still assigned to cluster = 1, eg. T03, T04 etc. Cluster = 2 and Cluster = 3 are mixed together. The K-means algorithm can determine k groups by defining the k value. We tested different nstart values and ran the program several times. The groups are maintained no matter how the nstart value changes. We can say that the kmeans method can determine which sampling sites belongs to which groups but we cannot extract the cluster types, in other words, we cannot say that the cluster = 1 that we generate by k-means is related to pure area or highland area.

The K-means algorithm determines the following three groups.

T03, T04, T05, T07, T11, T19, T63, T67
T01, T02, T06, T08, T10, T14, T16, T18, T23, T25, T34, T37, T39, T44, T51, T53, T56, T66, T68, T69, T70

Compared with CBH method with HTL training sort, we can see that K-means has a large overlap in group 2 and 3 when viewed through the convex hull m. As the
graph shows, it has only 8 sampling sites in group 1. However, in CBH HTL training, cluster = 1 has 18 sites. Although the CBH algorithm still has overlaps, the CBH Graham scan convex hull graph appears to be better than K-means(Figure 4.12).

Figure 4.11: NMS Plot: K-means & BJ Result

Figure 4.12: NMS Plot: CBH & K-means
4.9 Differences Between Three Algorithms

All three algorithms can get the cluster type, although the results are different. The biggest difference between these three algorithms is CBH can get the final corresponding cluster, but k-means, and PC-ord cannot. In k-means, the cluster type name would be changed for each run. Beijing results shows T2 to T8 belong to cluster = 2, but on the map, these sampling sites are located in the high lands area. Coincidentally, T26 to T35 shows cluster = 1(no or little pollution) in the Beijing results, but most of the sites are located in midlands area. CBH algorithm requires the original sampling sites location as a reference. Therefore, the results can reflect the extent of contamination at each of sampling sites.

There are very many parameters which will affect the CBH results, such as range, support, training percentage, training by different sort order. But in K-means, if we ignore the random cluster names, only the k value affects the cluster results.
Chapter 5

Conclusions and Future Work

5.1 Conclusions

In the thesis, we classify the Landuse data into three main cluster types based on watershed boundaries, 69 sampling sites in the river basin including 99 species. This $69 \times 99$ spare matrix includes 6831 cells, 5759 cells are empty. The apriori algorithm can work with character values. The use of individual mean values to determine the group range segments is a good way to represent the numeric data.

Using the apriori algorithm to find the association rules, we output the rules used to determine the cluster types. The highland (cluster = 1) has the most rules, and the lowland (cluster = 3) has the least rules. If we define Confidence = 0.1, and if other conditions are same, the confidence value does not affect the apriori results. However, support has a considerable effect. In this paper, we only consider Support = 0.04, 0.05 or 0.06. If the support value is increased, the rules tend to be reduced in number.

The training of these rules are organized using the CBH searching method. Training by association rules ordered by lift value from high to low or from low to high are both considered. User can change training percentages, 25% 50% and 75% of total
rules are the initial training percentage in the study. We test the rest of rules. Because high lift value’s rule order has higher similarity, training HTL could potentially get improved accuracy in training results, but test matching rates are not always satisfactory. However, training LTH training accuracy is lower than for the HTL method, but it can get a higher matching result.

Using training results to mark cells in the landuse dataset, and counting the species type of occupied cells (each species in different sampling sites), the result might be cluster = 1 or 2 or 3. If we calculate all species types in each sample site, the maximum population of species of a particular type has a significant impact on the sampling site cluster type. If the same value occurs in two types, the species representing a higher pollution area should determine the final cluster.

Comparing Beijing results created by commercial software, the CBH results appear to be an improvement. Cluster = 1 sites mostly located in mountain forest areas which are part of the highlands. However, in the Beijing original results, most samplings are located in hilly forest areas. Coincidentally, cluster = 2 (mid-polluted cluster type sampling in Beijing’ original results), are located in the highlands. It is a contradiction.

We can see this problem in the box plot. The CBH method shows the median value of farmland percentage catchment values are increased (from highland to lowland), but in the Beijing results, they go down first, and then increase sharply. Also, comparing the cluster type results, in range A, support =0.04, confidence = 0.1, 75% HTL training case, only 4 sampling sites show good (cluster =1), and 10 sites in CBH method show bad, but the Beijing results tell us those sites are good. It is a
contradiction. When we swap Beijing original results for cluster = 1 and cluster = 2 types. The results match better. Both “good” sites increase in number from 4 to 15, 0 sites in CBH show “good” but in Beijing the swapped result shows “bad”. The switched result is more reasonable. Additionally, in outlier analysis, CBH compared with Beijing swap results’ outlier are better than compared with original results.

Compared with K-means results, CBH method works well. K-means groups 40 out of 69 sampling sites as one group; it looks far different from the CBH and Beijing results.

Unlike K-means, the cluster type in apriori-CBH training method is not unique for a particular sampling station. Also, range selection, support and confident value setting, training rate and training from HTL or LTH lift value all affect the final cluster type for some sites.

The support vector machine (SVM) [23] approach was attempted as outlined in our thesis proposal but it did not yield satisfactory results.

In dominance analysis for farmland percentage catchment data, the turning point in the HTL method is close to the Beijing 2 results.

This was a small and sparse dataset, and had many quirks and overlaps. It forms a good, manageable study for attempting to reduce association rules into a logic base for determining the components of cluster analysis, and for adding new results as they come in.
5.2 Future work

This research was motivated by a specific problem of assessing environmental data in the Taizi River. It is not a general-purpose clustering tool, and it needs much more work, testing it using artificial data where the clusters are known. It did compare well with two other schemes on this problem which were outlined in our thesis proposal.

If we don’t know how many clusters are possible, this fact must be determined by other means. Also the amount of test data for analyzing the cluster rules on larger portions of a larger data set is unclear and has not been tested with enough extra data.

The data set is small compared to data in traditional data mining problems. Larger data sets must be used to test the formulation of the association rules portion. Parameters such as support and lift are not known in advance, so the experiments have to be performed to really see how useful it is. Our research only represents a modest beginning.
Bibliography


Appendix A

Training Results Files

R1-101s04c10p25AH.csv  R1-101s04c10p50AH.csv  R1-101s04c10p75AH.csv
R1-102s04c10p25AH.csv  R1-102s04c10p50AH.csv  R1-102s04c10p75AH.csv
R1-103s04c10p25AH.csv  R1-103s04c10p50AH.csv  R1-103s04c10p75AH.csv
R1-101s04c10p25AL.csv  R1-101s04c10p50AL.csv  R1-101s04c10p75AL.csv
R1-102s04c10p25AL.csv  R1-102s04c10p50AL.csv  R1-102s04c10p75AL.csv
R1-103s04c10p25AL.csv  R1-103s04c10p50AL.csv  R1-103s04c10p75AL.csv
R1-101s05c10p25AH.csv  R1-101s05c10p50AH.csv  R1-101s05c10p75AH.csv
R1-102s05c10p25AH.csv  R1-102s05c10p50AH.csv  R1-102s05c10p75AH.csv
R1-103s05c10p25AH.csv  R1-103s05c10p50AH.csv  R1-103s05c10p75AH.csv
R1-101s05c10p25AL.csv  R1-101s05c10p50AL.csv  R1-101s05c10p75AL.csv
R1-102s05c10p25AL.csv  R1-102s05c10p50AL.csv  R1-102s05c10p75AL.csv
R1-103s05c10p25AL.csv  R1-103s05c10p50AL.csv  R1-103s05c10p75AL.csv
R1-101s06c10p25AH.csv  R1-101s06c10p50AH.csv  R1-101s06c10p75AH.csv
R1-102s06c10p25AH.csv  R1-102s06c10p50AH.csv  R1-102s06c10p75AH.csv
R1-103s06c10p25AH.csv  R1-103s06c10p50AH.csv  R1-103s06c10p75AH.csv
R1-101s06c10p25AL.csv  R1-101s06c10p50AL.csv  R1-101s06c10p75AL.csv
R1-102s06c10p25AL.csv  R1-102s06c10p50AL.csv  R1-102s06c10p75AL.csv
R1-103s06c10p25AL.csv  R1-103s06c10p50AL.csv  R1-103s06c10p75AL.csv
## Appendix B

### Training Results Contents

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TrainResult-Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>00001cluster=IF001{sp111=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>00002cluster=IF001{sp092=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>00003cluster=IF001{sp075=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>00004cluster=IF001{sp045=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>00005cluster=IF001{sp078=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>00587cluster=IF001{sp081=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>00593cluster=IF001{sp140=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>00619cluster=IF001{sp045=1T2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>00620cluster=IF001{sp062=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>00016cluster=IF002{sp129=19T7775 sp136=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>00021cluster=IF002{sp079=19T7775 sp205=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>00023cluster=IF002{sp073=19T7775 sp205=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>00028cluster=IF002{sp005=1T2 sp073=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>00030cluster=IF002{sp032=19T7775 sp079=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>00051cluster=IF002{sp129=19T7775 sp132=1T2}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>00053cluster=IF002{sp079=19T7775 sp129=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>00599cluster=IF002{sp129=19T7775 sp199=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>00599cluster=IF002{sp129=19T7775 sp132=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>00607cluster=IF002{sp034=3T18 sp129=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>00608cluster=IF002{sp079=19T7775 sp097=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>00609cluster=IF002{sp097=19T7775 sp199=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>00617cluster=IF002{sp073=19T7775 sp129=19T7775}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>00172cluster=IF003{sp073=19T7775 sp132=1T2 sp199=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>00199cluster=IF003{sp073=19T7775 sp127=19T7775 sp199=3T18}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Training results: 101s05c10p75AH
<table>
<thead>
<tr>
<th>J</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TrainResult-Data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>00061cluster=1F002 (sp100=1T2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>00062cluster=1F001 (sp111=3T18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>00063cluster=1F001 (sp092=19T7775)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>00064cluster=1F001 (sp075=19T7775)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>00065cluster=1F001 (sp045=3T18)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>00066cluster=1F001 (sp078=19T7775)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>00067cluster=1F002 (sp052=1T2) sp073=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>00068cluster=1F002 (sp073=19T7775) sp096=1T2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>00069cluster=1F002 (sp132=19T7775) sp126=3T18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0006Accluster=1F002 (sp129=19T7775) sp127=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>00078cluster=1F002 (sp034=3T18) sp140=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>00080cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>00081cluster=1F002 (sp081=19T7775) sp140=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>00083cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>00084cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>00085cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>00086cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>00087cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>00088cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>00089cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>00090cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>00091cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>00092cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>00093cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>00094cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>00095cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>00096cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>00097cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>00098cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>00099cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>000A0cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>000A1cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>000A2cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>000A3cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>000A4cluster=1F002 (sp073=19T7775) sp081=19T7775</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Training results: 101s05c10p75AH
Appendix C

CBH Code Generation

```r
# Program: 03
# Objective: training
# Note: Generalization,
# Original/compare data comparedata/Original H after updated
# A   B   A V B
# A and B   B
# A or B
#

XSL=0
# XSL- number of item
DATA4S3<- read.csv('SJLS4.csv', head =TRUE)
DATA4CD<-nrow(DATA4S3)
DATAPA2S3<-read.csv('Parameter2.csv', head =TRUE)
XBL<-DATAPA2S3[1,4]
# XBL----training rate
XLSL<-round((DATA4CD*XBL)/100,0)
#XLSL----training amount

#Initialization
write.table(K0, file="SJS1.csv", append=FALSE, row=FALSE, col=FALSE)
write.table(DATA4S3, file="SJLS1.csv", append=TRUE, row=FALSE, col=FALSE)
write.table(K0, file="SJS3.csv", append=FALSE, row=FALSE, col=FALSE)
write.table(K0, file="SJS6.csv", append=FALSE, row=FALSE, col=FALSE)
write.table(K0, file="SJS9.csv", append=FALSE, row=FALSE, col=FALSE)
write.table(K0, file="SJLS10.csv", append=FALSE, row=FALSE, col=FALSE)

n16=XLSL+1
###LOOP 16 Start
# Objective:send data to SJLS10.csv training amount +1 to the end
###
while (n16<=DATA4CD){
  write.table(DATA4S3[n16,1], file= "SJLS10.csv", append=TRUE, row=FALSE, col=FALSE)
n16=n16+1
}
###LOOP 16 End

n17=1
###LOOP 17 Start
# Objective: select training data, use temp file SJLS3.csv
###
while (n17<=XLSL){
  write.table(DATA4S3[n17,1], file= "SJLS3.csv", append=TRUE, row=FALSE, col=FALSE)
n17=n17+1
}
###LOOP 17 End
```
DATA3J <- read.csv('SJLS3.csv', header=TRUE)
write.table(KSJ0, file="SJLS4.csv", append=FALSE, row=FALSE, col=FALSE)
write.table(DATA3J, file="SJLS4.csv", append=TRUE, row=FALSE, col=FALSE)
# SJLS4 save training data
write.table(KSJ0, file="SJLS3.csv", append=FALSE, row=FALSE, col=FALSE)
DATA4 <- read.csv('SJLS4.csv', header=TRUE)
DB4ZFC <- substr(DATA4[,1],1,1000)
write.table(DB4ZFC, file="SJLS6.csv", append=TRUE, row=FALSE, col=FALSE)

# Send 1st row data(SJLS4.csv) to SJLS6.csv as compare data
DATA6 <- read.csv('SJLS6.csv', header=TRUE)
row<-nrow(DATA6)

n18 = 1
##### LOOP 18 Start
# Objective: get data from SJLS4.csv
#####
while (n18 < row) {
  DB4XH <- substr(DATA4[n18,1],1,5)
  DB4XH <- array No.
  DB4ZFC <- substr(DATA4[n18,1],1,1000)
  # DB4ZFC -- array string
  D41 <- as.numeric(substr(DATA4[n18,1],16,18))
  # D41 is the data items amount in brackets
  D42 <- substr(DATA4[n18,1],20,(20+15*D41-1))
  # D42 compare content in SJLS4
  DATA6 <- read.csv('SJLS6.csv', header=TRUE)
  row= nrow(DATA6)
}

n19 = 1
##### LOOP 19 Start
# Objective: get SJLS6.csv data in each row
#####
while (n19 <= row) {
  DB6ZFC <- substr(DATA6[n19,1],1,1000)
  D61 <- as.numeric(substr(DATA6[n19,1],16,18))
  D61 <- 1
  # D61 number of items in bracket
  D62 <- substr(DATA6[n19,1],20,(20+15*D61-1))
  # D62
  if (D601 <= D41) XSL=D601 else XSL=D41
  # XSL items number, compare SJLS4 and SJLS6,
  # if same, select SJLS6 value.
  DBFHI=1
  # DBFHI --- compare matching
  TB = 1
  WB = 15

n20 = 1
##### LOOP 20 Start
# Objective: get data from SJLS4 in each group
#####
while (n20 <= D41) {
  D44 <- substr(D42,TB,WB)
  W1 = 1
  W2 = 15
n21=1
#
while (n21<=D61) {
    D64<-substr(D62,W1,W2)
    BJ1<-compare(D64,D44, equal=TRUE)
    if(isTRUE(BJ1)) DBFH1=DBFH1+1
    W1=W1+15
    W2=W2+15
    n21=n21+1
}
#

if (DBFH1==XSL) break
n19=n19+1
#

if (DBFH1==XSL) break
n18=n18+1
#

while (XSL>=1){
    if (XSL==DBFH1>=1)
        write.table(DB4ZFC, file="SJLS6.csv", append=TRUE, row=FALSE, col=FALSE)
        break
}

# Objectives: training(matching/not matching)
# Matching condition:
# 1. SJLS6 and SJLS6 all matching
# 2. All SJLS 6 data can find same data in SJLS4
# Not matching condition:
# 1. One row in SJLS4 VS data SJLS6, No same result
# 2. All SJLS 6 data can NOT find all same data in SJLS4
# NOTE: NOT matching, add SJLS4 row into SJLS6
# Matching, no change.
##
# n18=n18+1
##
##
##
##
DATA65J <- read.csv('SJLS6.csv', header=TRUE)
write.table(DATA65J, file='SJLS9.csv', append=TRUE, row=FALSE, col=FALSE)
DATA105J <- read.csv('SJLS1.csv', header=TRUE)
write.table(KS30, file='SJLS4.csv', append=FALSE, row=FALSE, col=FALSE)
write.table(DATASJ, file='SJLS4.csv', append=TRUE, row=FALSE, col=FALSE)

XLBLZF <- format(XLBL)
# XLBLZF Training %

NEWFILE1 <- paste('Train', DATANAME, OUTFILEWORD1, sep='", collapse=NULL)
DATA9S <- read.csv('SJLS9.csv', header=FALSE)
write.table(DATA9S, file=NEWFILE1, append=FALSE, row=FALSE, col=FALSE)
# sent result to New file

NEWFILE2 <- paste('Test', DATANAME, OUTFILEWORD1, sep='", collapse=NULL)
# NEWFILE2 new file
DATA105S <- read.csv('SJLS10.csv', header=FALSE)
write.table(DATA105S, file=NEWFILE2, append=FALSE, row=FALSE, col=FALSE)
# Test data send to New file 2

NEWFILE3 <- paste('R0-', DATANAME, OUTFILEWORD1, sep='", collapse=NULL)
setwd(WORKPATH4)
# WORKPATH4 - results path
write.table('Train-Data', file=NEWFILE3, append=FALSE, row=FALSE, col=FALSE)
write.table(TRAINNDATA, file=NEWFILE3, append=TRUE, row=FALSE, col=FALSE)
setwd(WORKPATH2)
write.table(NEWFILE1, file='SJLS7.csv', append=TRUE, row=FALSE, col=FALSE)
write.table(NEWFILE2, file='SJLS8.csv', append=TRUE, row=FALSE, col=FALSE)

#### LOOP 23 Start
# Objective: after training, create training result file name
# add to Parameter2
####
while (n1 >= 3){
# n1 loop start at 01 program, and end at 03 program
DATA7 <- read.csv('SJLS7.csv', header=TRUE)
DATA7LZH <- t(DATA7)
# DATA7LZH col to row
write.table(DATA7LZH, file='Parameter2.csv', append=TRUE, row=FALSE, col=FALSE, sep='","
DATA8 <- read.csv('SJLS8.csv', header=TRUE)
DATA8LZH <- t(DATA8)
# DATA8LZH col to row
write.table(DATA8LZH, file='Parameter2.csv', append=TRUE, row=FALSE, col=FALSE, sep='","
break}
#### LOOP 23 End

n1 = n1+1
}
#### LOOP 1 End
setwd(WORKPATH1)
# WORKPATH1