# Data Mining and Visual Analytics on Three Gorges Project

Adithya Addanki
Department of Computer Science,
The University of Akron,
Akron, Ohio-44304, *aa207@zips.uakron.edu*

Deekshith Sandesari
Department of Computer Science,
The University of Akron,
Akron, Ohio-44304, *ds168@zips.uakron.edu*

Srinivas Rao Katta
Department of Computer Science,
The University of Akron,
Akron, Ohio-44304, *sk189@zips.uakron.edu*

Lakshmi Sharnitha Yarlagadda
Department of Computer Science,
The University of Akron,
Akron, Ohio-44304, *ly27@zips.uakron.edu*

## Abstract

*Growth of a country lies not just in the development of assets or infrastructures but also with the wellbeing of the country's population. There has always been issues with involuntary migration with regards to various aspects of the people pre and post relocations. In this paper we try to analyze the different dominant characteristics of the people subject to migration, through a panel study data collected during the construction of Three Gorges Dam. We apply Knowledge Discovery processes to mine the dominant attributes in determining the status of migrants and non-migrants. The paper also discusses the tools we have developed to visualize the patterns in the data. These visualizers have been developed as prototypes to better understand the data, their distributions and other aspects which we discuss in the following sections. We conclude the paper with the possible extensions to the work in related domains so as to give a better perspective to improve the country's development without deteriorating the lives of people affected because of the relocation.*

## 1. Introduction

The primary goal of this project is to analyze the dataset collected through a Prospective panel study involving two phases(late 2002 to early 2003 and 2006) of face to face interviews with 1530 people(975 designated migrants and 555 non-migrants). This data set has been used for various analyses on the statuses of the migrants, which reflected the physical and mental health conditions as a result of involuntary migration.

## 2. Background Study

This section presents a brief idea of the concepts and terminology we would use for the rest of the sections in the report.

### 2.1 Data Mining

Data Mining refers to the application of algorithms for extracting patterns from data. Different tasks that could be performed through data mining include classification (identifying classification rules), clustering, finding association rules etc. A data mining system typically has the ability to generate thousand or even millions of rules but not all of them are interesting. A pattern is interesting if it leads to revelation of an unknown fact.

#### 2.1.1 KDD

Knowledge Discovery in Data refers to the overall process of finding and interpreting useful patterns in data. With ever growing data, we are drowning in information but relevant knowledge is still below par. With the application of KDD we find relevant information that helps us understand the data better and make proper and effective use of the mined information. It has found its applications in various areas including but not limited to Market Basket analysis, catalog design, store layout, customer satisfaction, patient’s illness diagnosis, processing loan & credit card applications, etc.

### 2.2 Weka

Weka is an open source software issued under the GNU General Public License. Weka is a collection of machine learning algorithms for data mining tasks [1]. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka allows application of algorithms it offers directly to a dataset or interface it to custom developed Java code through its API. Documentation and other resources for Weka Application could be found at the official site of Weka [2].

### 2.3 Microsoft SQL Server

Microsoft SQL server is a relational database management system offering functionality that stores and retrieves data as required. Microsoft SQL Server offers breakthrough in-memory performance built into the database for your transactions, your queries and your analytics, delivers faster insights into data with familiar analytic tools [3].

### 2.4 Visualization

Visual Communication is the fundamental process for exploring and dissemination of information through visual aids such as diagrams, sketches, photographs, charts, video, animations etc., thus improving the quality of comprehension, memory and inference. The strength of visualizations lies in the design principles which connect the visualizations with viewer's perception and cognition of underlying information that is conveyed [4].

#### 2.4.1 Parallel Coordinates

Visualizing information that has more than 3 dimensions is not understandable to the human eye. Parallel coordinates are introduced by Philbert Maurice D'Ocagne in 1885[5]. Parallel coordinates are one of the most famous visualization techniques, while initially confusing, they are a very powerful tool for understanding multi-dimensional numerical datasets [6]. Parallel Coordinates represent multiple dimensions with the help of parallel axes and samples in data set are represented as poly lines spanning the axes.

### 2.4.2 Interactive Visualizer using Bubble Chart

Interactive Visualization allow the data analyst to directly interact with the visualizations, dynamically change the constructs for further analysis and exploration. It is often found easy to relate and combine multiple independent visualizations [9]. Bubble Charts (3 Dimensions) is one such technique where two dimensions are visualized as co-ordinate axes and other as size of the bubble [10]. Filtering mechanism helps in partitioning the data into subsets and focus on interesting subjects/patterns [8].

### 2.4.3 Geo-Visualization using Google Maps API

Geo-visualization is the geospatial analysis of data by using interactive visualization techniques. Geo-Visualization usually deals with creating static or dynamic interactive maps that has customized markers and polylines to depict relevant information with respect to the locations on the map.

## 3. Data set Description

Three gorges project dataset has recorded responses to a questionnaire before and after the relocation of 1.4 million Chinese people. Though started with 1530 subjects, the second wave was able to capture a success rate of 70%, thus resulting in the final sample size of 1056. Of these 1056 subjects; 350 are Non-Migrants, 286 are designated Migrants who have not yet moved, 420 remaining subjects who are designated and moved. The dataset was collected with the results from the samples for the questionnaire used for the survey that covered various aspects of the subject’s life. A short description of the dataset is given below.

Attributes in the data could be grouped into the following categories:

- Health: Mental and Physical Health
- Psychological Resources: Sense of Control, Self-Esteem
- Social Resources: Perceived Routine Support, Social Connections and Social Comparison
- Relocation Status
- Demographic Variables
- Attitudes towards relocation
- Integration with Local Community
- Compensation
- Employment, Income, Health Life Style and Housing

Questionnaire sample for health related attributes is given below:

- How often during the last week 1) were you bothered by things that usually don't bother you? 2) How often have you felt like everything you did was an effort? 3) How often have you felt that you were just as good as other people*? 4) How often have you had trouble keeping your mind on what you were doing? ….
- Responses are coded 0 for less than once a week, 1 for one to two days a week, 2 for three to four days a week, 3 for five to seven days a week

## 4. Data Mining Approach

This section describes the procedure followed for mining the dataset in order to observe the characteristics of the relocated people pre and post migration. To get accurate consequences of the forced migration, the approach is split into three case to perform the mining and KDD process.

Case1: Phase 1[i.e. pre-migration dataset],

Case2: Phase 2[i.e. Post migration dataset] and

Case3: Common Dataset [i.e. Common attributes dataset of Phase1 and Phase2].

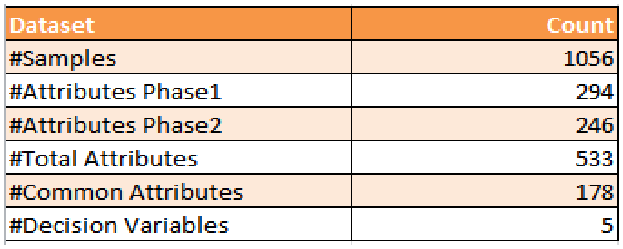
The further steps in our approach are to handle the missing values, analysis, imputation oriented analysis based on the suggestions from domain experts and the analysis of the common attributes in both the phases. The detailed explanation for every single step follows.

## 4.1 Data Collection

The Dataset is a copy righted source collected with authorization from Dr. Juan Xi, Associate Professor of Sociology at The University of Akron. The study involves two phases of interviews (pre-migration and post-migration).

## 4.2 Data Preprocessing

Data pre-processing is an important step in Data Mining as it handles noisy and incomplete data. The original dataset has all the responses coded in a single file. For our analysis we had to split the data set into two sets; Phase 1/Wave1 (pre-migration) and Phase 2/ Wave2 (post-migration), so as to study the two phases individually.



| Dataset | Count |
|---|---|
| #Samples | 1056 |
| #Attributes Phase1 | 294 |
| #Attributes Phase2 | 246 |
| #Total Attributes | 533 |
| #Common Attributes | 178 |
| #Decision Variables | 5 |

*Table 4.2.1: Total Number of Attributes in Phase1 and Phase2*

Data is preprocessed by handling the missing values, handling the missing values in dependent attributes, and variance in the attributes of two phases. For Instance: the variance of the attributes for employment category, 7 attributes in phase1 and 5 attributes in phase2 and the similarly income category has 65 attributes in phase1 and 42 attributes in phase2.

### 4.2.1 Data Cleaning

#### 4.2.1.1 Handling Missing Values Related to Dependent Attributes

In our data set, compensation related attributes of non-migrants are missing, which in reality is not a required value for analysis. Compensation is decided based on the discussions with higher officials, the person may or may not be satisfied, which is a dependent variable. When it is not required or in case of non-migrants assumed to be zero based on Domain expert's suggestion.

#### 4.2.1.2. Variance in Attributes of Two Phases

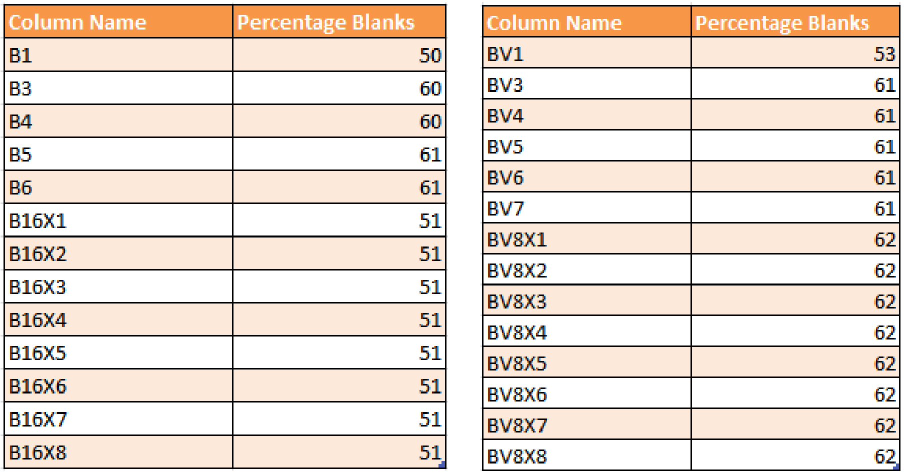
To handle the variance in attributes, the data is split into two phases (Phase1, Phase2) initially and then the results are analyzed in two perspectives, one with all the attributes in Phase1 and Phase2 and the other with only common attributes in Phase1 and Phase2.

#### 4.2.1.3. Ignoring the Attributes

In the dataset, there are three attributes related to the location address of each sample which are deleted as these were encoded in Chinese and are treated as irrelevant for our study.

#### 4.2.1.4 Handling the Missing Values

The dataset is loaded into data preprocessing step in WEKA Explorer to identify the missing attribute values and the missing percentage ratio for each attribute. These missing values and missing percentage ratio are validated by loading the data into Microsoft SQL Server 2012. Table2 and Table3 demonstrates the sample attributes in Phase1 and Phase2 containing missing values.

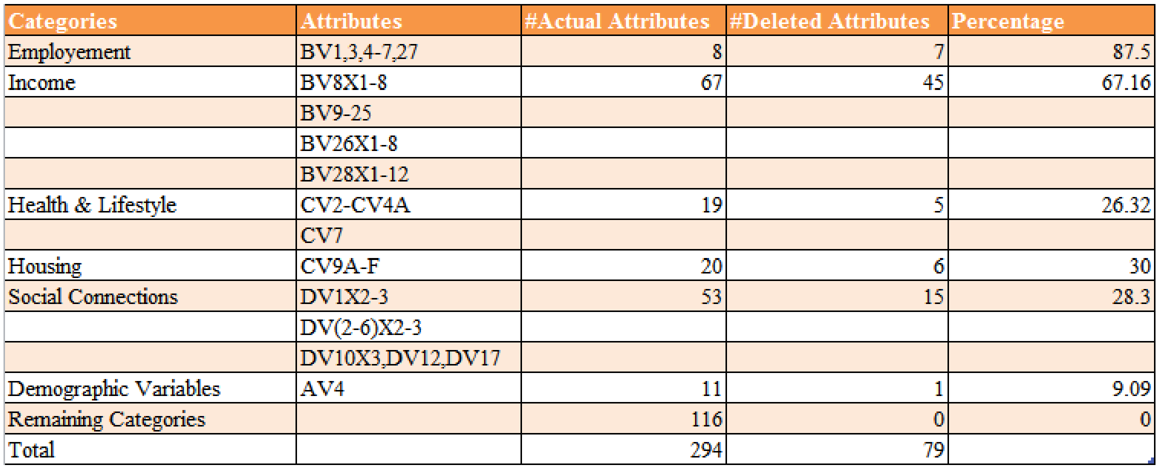


| Column Name | Percentage Blanks |
|---|---|
| B1 | 50 |
| B3 | 60 |
| B4 | 60 |
| B5 | 61 |
| B6 | 61 |
| B16X1 | 51 |
| B16X2 | 51 |
| B16X3 | 51 |
| B16X4 | 51 |
| B16X5 | 51 |
| B16X6 | 51 |
| B16X7 | 51 |
| B16X8 | 51 |

| Column Name | Percentage Blanks |
|---|---|
| BV1 | 53 |
| BV3 | 61 |
| BV4 | 61 |
| BV5 | 61 |
| BV6 | 61 |
| BV7 | 61 |
| BV8X1 | 62 |
| BV8X2 | 62 |
| BV8X3 | 62 |
| BV8X4 | 62 |
| BV8X5 | 62 |
| BV8X6 | 62 |
| BV8X7 | 62 |
| BV8X8 | 62 |

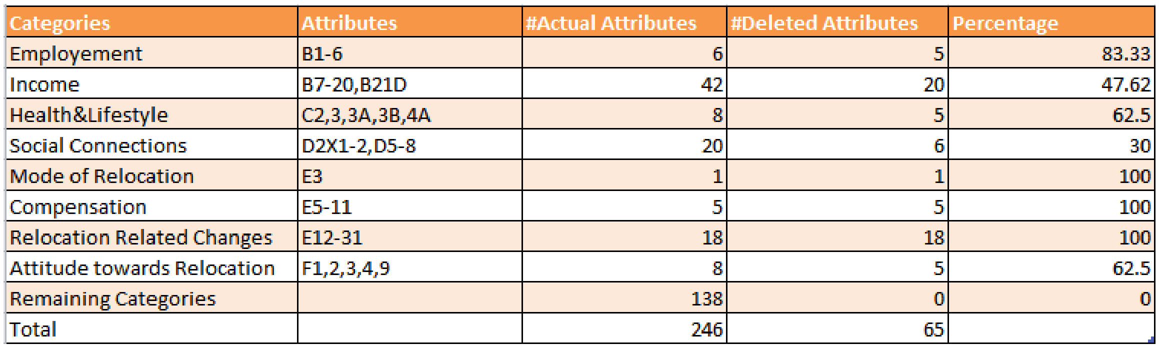
*Table 4.2.1.4.1: Sample Missing attributes and their Percentages (Phase 1-Left and Phase2-Right)*

The original dataset which is being split into two phases are loaded into Microsoft SQL Server 2012 to delete the attributes whose missing percentage ratio is greater than 30%. Table 4.2.1.4.2 and Table 4.2.1.4.3 gives detailed statistical results of missing values and its percentage in each category for Phase1 and Phase2.



| Categories | Attributes | #Actual Attributes | #Deleted Attributes | Percentage |
|---|---|---|---|---|
| Employement | BV1,3,4-7,27 | 8 | 7 | 87.5 |
| Income | BV8X1-8 | 67 | 45 | 67.16 |
| | BV9-25 | | | |
| | BV26X1-8 | | | |
| | BV28X1-12 | | | |
| Health & Lifestyle | CV2-CV4A | 19 | 5 | 26.32 |
| | CV7 | | | |
| Housing | CV9A-F | 20 | 6 | 30 |
| Social Connections | DV1X2-3 | 53 | 15 | 28.3 |
| | DV(2-6)X2-3 | | | |
| | DV10X3,DV12,DV17 | | | |
| Demographic Variables | AV4 | 11 | 1 | 9.09 |
| Remaining Categories | | 116 | 0 | 0 |
| Total | | 294 | 79 | |

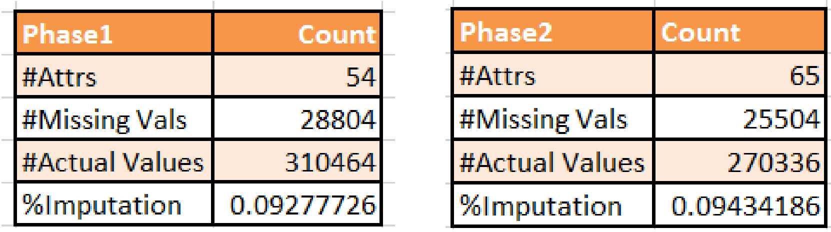
*Table 4.2.1.4.2: Phase1 Categories containing missing attributes and their percentage ratio*



| Categories | Attributes | #Actual Attributes | #Deleted Attributes | Percentage |
|---|---|---|---|---|
| Employement | B1-6 | 6 | 5 | 83.33 |
| Income | B7-20,B21D | 42 | 20 | 47.62 |
| Health&Lifestyle | C2,3,3A,3B,4A | 8 | 5 | 62.5 |
| Social Connections | D2X1-2,D5-8 | 20 | 6 | 30 |
| Mode of Relocation | E3 | 1 | 1 | 100 |
| Compensation | E5-11 | 5 | 5 | 100 |
| Relocation Related Changes | E12-31 | 18 | 18 | 100 |
| Attitude towards Relocation | F1,2,3,4,9 | 8 | 5 | 62.5 |
| Remaining Categories | | 138 | 0 | 0 |
| Total | | 246 | 65 | |

*Table 4.2.1.4.3: Phase2 Categories containing missing attributes and their percentage ratio*

The attributes in each category for Phase1 and Phase 2 containing greater than 30% missing values are deleted and with the help of SQL Server 2012, developed a PL/SQL script to replace the blank and white spaced remaining attributes whose missing percentage values less than 30% is imputed with "?". Table 4.2.1.4.4 displays the Imputation statistics for Phase1 and Phase2

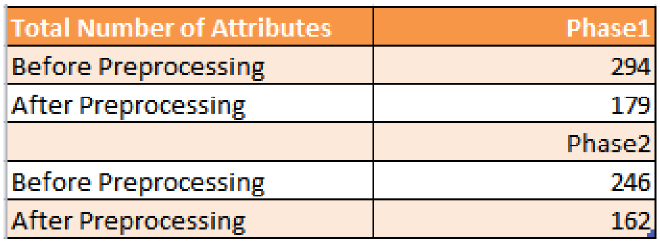


| Phase1 | Count |
|---|---|
| #Attrs | 54 |
| #Missing Vals | 28804 |
| #Actual Values | 310464 |
| %Imputation | 0.09277726 |

| Phase2 | Count |
|---|---|
| #Attrs | 65 |
| #Missing Vals | 25504 |
| #Actual Values | 270336 |
| %Imputation | 0.09434186 |

*Table 4.2.1.4.4: Imputation Statistics for Phase1 (Left) and Phase2 (Right)*

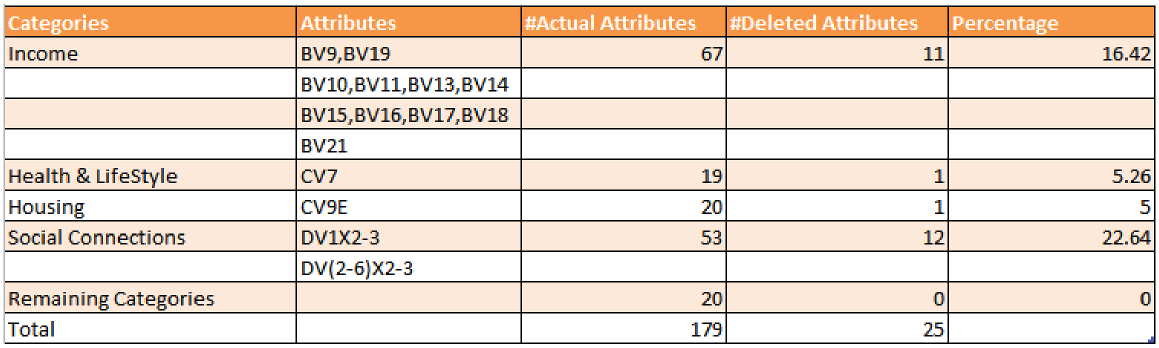
Table 4.2.1.4.5 displays the total number of attributes present in Phase1 and Phase2, after deletion of the attributes with greater than 30% missing values.



| Total Number of Attributes | Phase1 |
|---|---|
| Before Preprocessing | 294 |
| After Preprocessing | 179 |
| | Phase2 |
| Before Preprocessing | 246 |
| After Preprocessing | 162 |

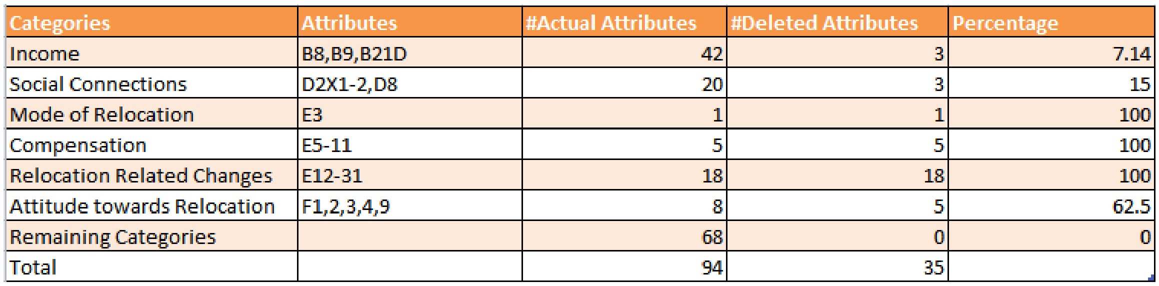
*Table 4.2.1.4.5: Total Number of attributes before preprocessing and after pre-processing in Phase1 and Phase2*

The next step of data cleaning is to load the original dataset which is split into two phases again and impute the attribute values which are dependent on other attributes but the values are missing. The dependent attribute needs to be imputed with the values given by domain expert so that results will not be biased. Example of a dependent attribute imputation: when BV1=3 and BV3='?' then BV3 is updated to 0(introducing a new category into the attribute BV3). After imputation of attributes with values and the truly missing values with "?", Table 4.2.1.4.6 and Table 4.2.1.4.7 describes in details the statistics of missing attributes and its percentage for each category for Phase1 and Phase2.



| Categories | Attributes | #Actual Attributes | #Deleted Attributes | Percentage |
|---|---|---|---|---|
| Income | BV9,BV19 | 67 | 11 | 16.42 |
| | BV10,BV11,BV13,BV14 | | | |
| | BV15,BV16,BV17,BV18 | | | |
| | BV21 | | | |
| Health & LifeStyle | CV7 | 19 | 1 | 5.26 |
| Housing | CV9E | 20 | 1 | 5 |
| Social Connections | DV1X2-3 | 53 | 12 | 22.64 |
| | DV(2-6)X2-3 | | | |
| Remaining Categories | | 20 | 0 | 0 |
| Total | | 179 | 25 | |

*Table 4.2.1.4.6: Statistics for number of missing attributes after imputation for Phase1*



| Categories | Attributes | #Actual Attributes | #Deleted Attributes | Percentage |
|---|---|---|---|---|
| Income | B8,B9,B21D | 42 | 3 | 7.14 |
| Social Connections | D2X1-2,D8 | 20 | 3 | 15 |
| Mode of Relocation | E3 | 1 | 1 | 100 |
| Compensation | E5-11 | 5 | 5 | 100 |
| Relocation Related Changes | E12-31 | 18 | 18 | 100 |
| Attitude towards Relocation | F1,2,3,4,9 | 8 | 5 | 62.5 |
| Remaining Categories | | 68 | 0 | 0 |
| Total | | 94 | 35 | |

*Table 4.2.1.4.7: Statistics for number of missing attributes after imputation for Phase2*

As the final step of the data cleaning process, all the attributes having a missing percentage ratio greater than 30% are deleted. With the help of Weka and SQL Server, the data is cleaned by replacing certain attributes of interest with values, deleting missing attributes by considering the missing threshold as 30% and ignoring few attributes.

## 4.3 Data Transformation

With the data cleaning process in various cases, we have reduced the data set in terms of attributes which have percentage of missing values greater than 30%. The other case where we have transformed the data; identified the common attributes in both the phases and renamed them to follow a naming convention. For instance Employment related attributes are renamed as EMPx (where x is the number of attributes that matched in that group). Thus transforming the original data set in both phases to contain only the common attributes for one of our analysis.

## 4.4 Data Mining

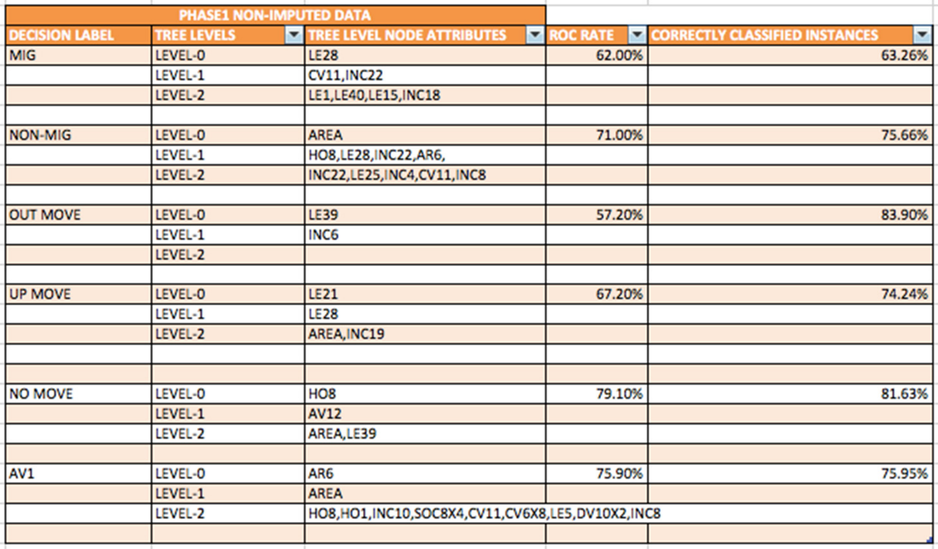
The data mining task at hand is classification as we like to determine the dominant attributes that help in classifying the respondents of the survey to be a migrant or non-migrant.

### 4.4.1 Classification Models

To build a classifier model we used J48 classifier with 10-fold cross validation using Weka tool to generate a decision tree. We have generated classification models for the sets of data that have been renamed in data transformation stage and updated with missing values with '?' in data cleaning stage and deleted the attributes that have greater that 30% missing values in data reduction stage. The decision attributes considered are migrant, non-migrant, no-move, up-move and out-move. The classification models are outlined below:

*Case1: For all Phase1 attributes*

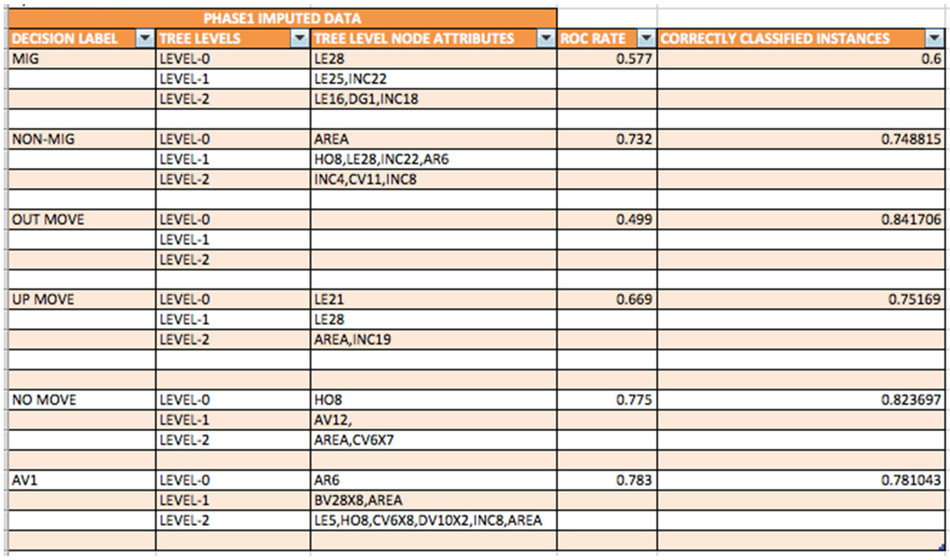
a) *Without Imputation:* We have considered the Phase1 attributes, performed the classification and reached at the following results [Table 4.4.1.1].



| PHASE1 NON-IMPUTED DATA | | | | |
|---|---|---|---|---|
| DECISION LABEL | TREE LEVELS | TREE LEVEL NODE ATTRIBUTES | ROC RATE | CORRECTLY CLASSIFIED INSTANCES |
| MIG | LEVEL-0 | LE28 | 62.00% | 63.26% |
| | LEVEL-1 | CV11,INC22 | | |
| | LEVEL-2 | LE1,LE40,LE15,INC18 | | |
| | | | | |
| NON-MIG | LEVEL-0 | AREA | 71.00% | 75.66% |
| | LEVEL-1 | HO8,LE28,INC22,AR6, | | |
| | LEVEL-2 | INC22,LE25,INC4,CV11,INC8 | | |
| | | | | |
| OUT MOVE | LEVEL-0 | LE39 | 57.20% | 83.90% |
| | LEVEL-1 | INC6 | | |
| | LEVEL-2 | | | |
| | | | | |
| UP MOVE | LEVEL-0 | LE21 | 67.20% | 74.24% |
| | LEVEL-1 | LE28 | | |
| | LEVEL-2 | AREA,INC19 | | |
| | | | | |
| | | | | |
| NO MOVE | LEVEL-0 | HO8 | 79.10% | 81.63% |
| | LEVEL-1 | AV12 | | |
| | LEVEL-2 | AREA,LE39 | | |
| | | | | |
| AV1 | LEVEL-0 | AR6 | 75.90% | 75.95% |
| | LEVEL-1 | AREA | | |
| | LEVEL-2 | HO8,HO1,INC10,SOC8X4,CV11,CV6X8,LE5,DV10X2,INC8 | | |
| | | | | |

*Table 4.4.1.1. Phase1 attributes without imputation*

b) *With Imputation:* We have done the imputation on the above data of phase1 by considering the suggestions from Domain expert, performed the classification and reached at the following results [Table 4.4.1.2].

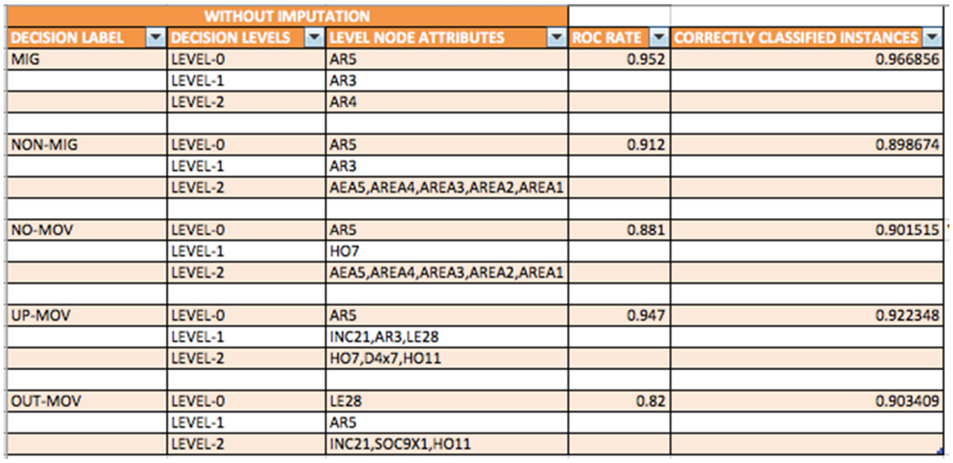


| PHASE1 IMPUTED DATA | | | | |
|---|---|---|---|---|
| DECISION LABEL | TREE LEVELS | TREE LEVEL NODE ATTRIBUTES | ROC RATE | CORRECTLY CLASSIFIED INSTANCES |
| MIG | LEVEL-0 | LE28 | 0.577 | 0.6 |
| | LEVEL-1 | LE25,INC22 | | |
| | LEVEL-2 | LE16,DG1,INC18 | | |
| | | | | |
| NON-MIG | LEVEL-0 | AREA | 0.732 | 0.748815 |
| | LEVEL-1 | HO8,LE28,INC22,AR6 | | |
| | LEVEL-2 | INC4,CV11,INC8 | | |
| | | | | |
| OUT MOVE | LEVEL-0 | | 0.499 | 0.841706 |
| | LEVEL-1 | | | |
| | LEVEL-2 | | | |
| | | | | |
| UP MOVE | LEVEL-0 | LE21 | 0.669 | 0.75169 |
| | LEVEL-1 | LE28 | | |
| | LEVEL-2 | AREA,INC19 | | |
| | | | | |
| | | | | |
| NO MOVE | LEVEL-0 | HO8 | 0.775 | 0.823697 |
| | LEVEL-1 | AV12, | | |
| | LEVEL-2 | AREA,CV6X7 | | |
| | | | | |
| AV1 | LEVEL-0 | AR6 | 0.783 | 0.781043 |
| | LEVEL-1 | BV28X8,AREA | | |
| | LEVEL-2 | LE5,HO8,CV6X8,DV10X2,INC8,AREA | | |
| | | | | |

*Table 4.4.1.2. Phase1 attributes with imputation*

*Case2: For all Phase2 attributes*

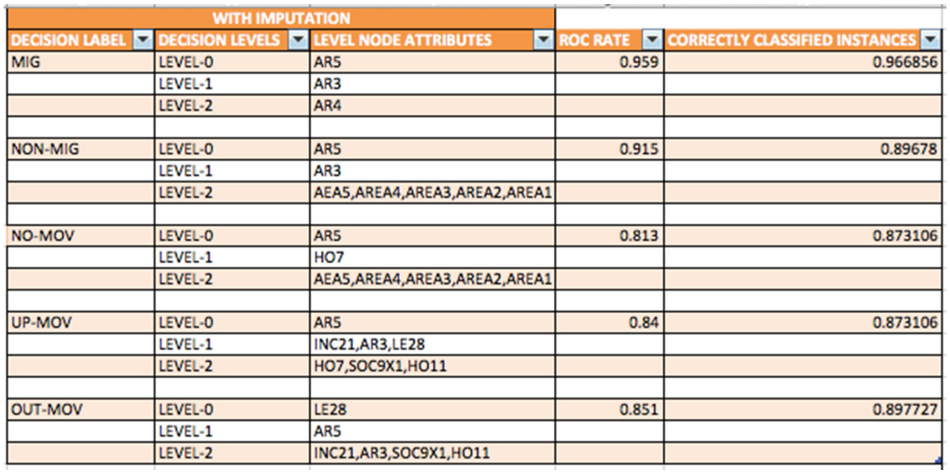
a) *Without Imputation:* We have considered the Phase2 attributes, performed the classification and reached at the following results [Table 4.4.1.3].



| WITHOUT IMPUTATION | | | | |
|---|---|---|---|---|
| DECISION LABEL | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | CORRECTLY CLASSIFIED INSTANCES |
| MIG | LEVEL-0 | AR5 | 0.952 | 0.966856 |
| | LEVEL-1 | AR3 | | |
| | LEVEL-2 | AR4 | | |
| | | | | |
| NON-MIG | LEVEL-0 | AR5 | 0.912 | 0.898674 |
| | LEVEL-1 | AR3 | | |
| | LEVEL-2 | AEA5,AREA4,AREA3,AREA2,AREA1 | | |
| | | | | |
| NO-MOV | LEVEL-0 | AR5 | 0.881 | 0.901515 |
| | LEVEL-1 | HO7 | | |
| | LEVEL-2 | AEA5,AREA4,AREA3,AREA2,AREA1 | | |
| | | | | |
| UP-MOV | LEVEL-0 | AR5 | 0.947 | 0.922348 |
| | LEVEL-1 | INC21,AR3,LE28 | | |
| | LEVEL-2 | HO7,D4x7,HO11 | | |
| | | | | |
| OUT-MOV | LEVEL-0 | LE28 | 0.82 | 0.903409 |
| | LEVEL-1 | AR5 | | |
| | LEVEL-2 | INC21,SOC9X1,HO11 | | |

*Table 4.4.1.3. Phase2 attributes without imputation*

b) *With Imputation:* We have done the imputation on the above data related to phase2 by considering the suggestions from Domain expert, performed the classification and reached at the following results [Table 4.4.1.4].



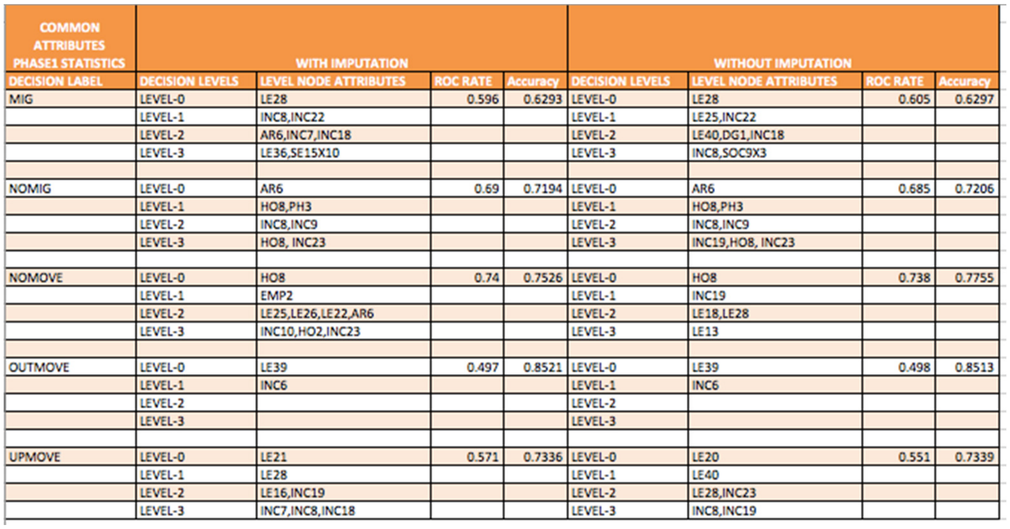
| WITH IMPUTATION | | | | |
|---|---|---|---|---|
| DECISION LABEL | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | CORRECTLY CLASSIFIED INSTANCES |
| MIG | LEVEL-0 | AR5 | 0.959 | 0.966856 |
| | LEVEL-1 | AR3 | | |
| | LEVEL-2 | AR4 | | |
| | | | | |
| NON-MIG | LEVEL-0 | AR5 | 0.915 | 0.89678 |
| | LEVEL-1 | AR3 | | |
| | LEVEL-2 | AEA5,AREA4,AREA3,AREA2,AREA1 | | |
| | | | | |
| NO-MOV | LEVEL-0 | AR5 | 0.813 | 0.873106 |
| | LEVEL-1 | HO7 | | |
| | LEVEL-2 | AEA5,AREA4,AREA3,AREA2,AREA1 | | |
| | | | | |
| UP-MOV | LEVEL-0 | AR5 | 0.84 | 0.873106 |
| | LEVEL-1 | INC21,AR3,LE28 | | |
| | LEVEL-2 | HO7,SOC9X1,HO11 | | |
| | | | | |
| OUT-MOV | LEVEL-0 | LE28 | 0.851 | 0.897727 |
| | LEVEL-1 | AR5 | | |
| | LEVEL-2 | INC21,AR3,SOC9X1,HO11 | | |

*Table 4.4.1.4. Phase2 attributes with imputation*

*Case3a: Common attributes in Phase1*

a) *Without Imputation:* We have considered only common attributes in Phase1, performed the classification and reached at the results [Table 4.4.1.5 Right].

b) *With Imputation:* We have done the imputation on the above data related to common attributes in Phase1 by considering the suggestions from Domain expert, performed the classification and reached at the results [Table 4.4.1.5 Left].



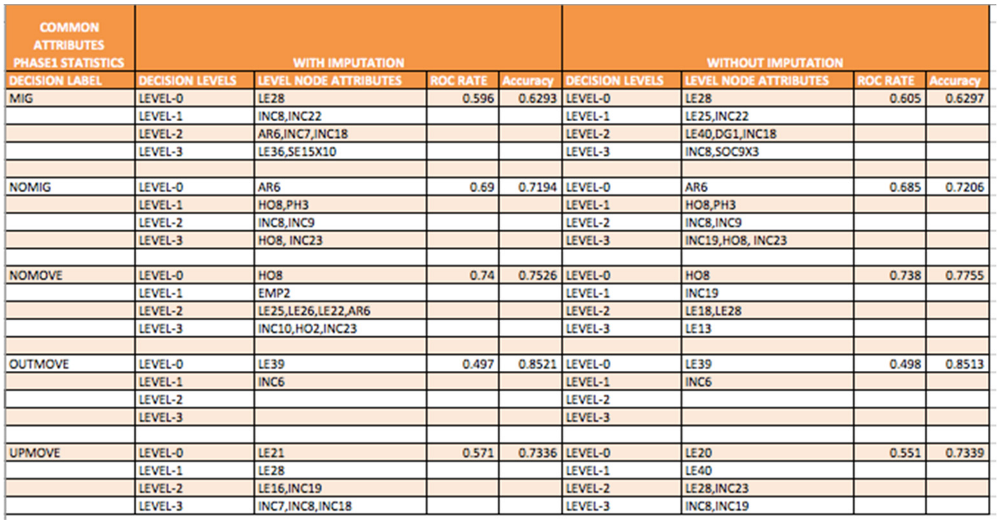
| COMMON ATTRIBUTES PHASE1 STATISTICS | WITH IMPUTATION | | | | WITHOUT IMPUTATION | | | |
|---|---|---|---|---|---|---|---|---|
| DECISION LABEL | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | Accuracy | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | Accuracy |
| MIG | LEVEL-0 | LE28 | 0.596 | 0.6293 | LEVEL-0 | LE28 | 0.605 | 0.6297 |
| | LEVEL-1 | INC8,INC22 | | | LEVEL-1 | LE25,INC22 | | |
| | LEVEL-2 | AR6,INC7,INC18 | | | LEVEL-2 | LE40,DG1,INC18 | | |
| | LEVEL-3 | LE36,SE15X10 | | | LEVEL-3 | INC8,SOC9X3 | | |
| | | | | | | | | |
| NOMIG | LEVEL-0 | AR6 | 0.69 | 0.7194 | LEVEL-0 | AR6 | 0.685 | 0.7206 |
| | LEVEL-1 | HO8,PH3 | | | LEVEL-1 | HO8,PH3 | | |
| | LEVEL-2 | INC8,INC9 | | | LEVEL-2 | INC8,INC9 | | |
| | LEVEL-3 | HO8, INC23 | | | LEVEL-3 | INC19,HO8, INC23 | | |
| | | | | | | | | |
| NOMOVE | LEVEL-0 | HO8 | 0.74 | 0.7526 | LEVEL-0 | HO8 | 0.738 | 0.7755 |
| | LEVEL-1 | EMP2 | | | LEVEL-1 | INC19 | | |
| | LEVEL-2 | LE25,LE26,LE22,AR6 | | | LEVEL-2 | LE18,LE28 | | |
| | LEVEL-3 | INC10,HO2,INC23 | | | LEVEL-3 | LE13 | | |
| | | | | | | | | |
| OUTMOVE | LEVEL-0 | LE39 | 0.497 | 0.8521 | LEVEL-0 | LE39 | 0.498 | 0.8513 |
| | LEVEL-1 | INC6 | | | LEVEL-1 | INC6 | | |
| | LEVEL-2 | | | | LEVEL-2 | | | |
| | LEVEL-3 | | | | LEVEL-3 | | | |
| | | | | | | | | |
| UPMOVE | LEVEL-0 | LE21 | 0.571 | 0.7336 | LEVEL-0 | LE20 | 0.551 | 0.7339 |
| | LEVEL-1 | LE28 | | | LEVEL-1 | LE40 | | |
| | LEVEL-2 | LE16,INC19 | | | LEVEL-2 | LE28,INC23 | | |
| | LEVEL-3 | INC7,INC8,INC18 | | | LEVEL-3 | INC8,INC19 | | |

*Table 4.4.1.5. Phase1 common attributes with and without imputation*

*Case3b: Common attributes in Phase2*

c) *Without Imputation:* We have considered only the common attributes in Phase2, performed the classification and reached at the results [Table 4.4.1.6 Right].

d) *With Imputation:* We have done the imputation on the above data related to common attributes in Phase2 by considering the suggestions from Domain expert, performed the classification and reached at the results [Table 4.4.1.6 Left].



| COMMON ATTRIBUTES PHASE1 STATISTICS | WITH IMPUTATION | | | | WITHOUT IMPUTATION | | | |
|---|---|---|---|---|---|---|---|---|
| DECISION LABEL | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | Accuracy | DECISION LEVELS | LEVEL NODE ATTRIBUTES | ROC RATE | Accuracy |
| MIG | LEVEL-0 | LE28 | 0.596 | 0.6293 | LEVEL-0 | LE28 | 0.605 | 0.6297 |
| | LEVEL-1 | INC8,INC22 | | | LEVEL-1 | LE25,INC22 | | |
| | LEVEL-2 | AR6,INC7,INC18 | | | LEVEL-2 | LE40,DG1,INC18 | | |
| | LEVEL-3 | LE36,SE15X10 | | | LEVEL-3 | INC8,SOC9X3 | | |
| | | | | | | | | |
| NOMIG | LEVEL-0 | AR6 | 0.69 | 0.7194 | LEVEL-0 | AR6 | 0.685 | 0.7206 |
| | LEVEL-1 | HO8,PH3 | | | LEVEL-1 | HO8,PH3 | | |
| | LEVEL-2 | INC8,INC9 | | | LEVEL-2 | INC8,INC9 | | |
| | LEVEL-3 | HO8, INC23 | | | LEVEL-3 | INC19,HO8, INC23 | | |
| | | | | | | | | |
| NOMOVE | LEVEL-0 | HO8 | 0.74 | 0.7526 | LEVEL-0 | HO8 | 0.738 | 0.7755 |
| | LEVEL-1 | EMP2 | | | LEVEL-1 | INC19 | | |
| | LEVEL-2 | LE25,LE26,LE22,AR6 | | | LEVEL-2 | LE18,LE28 | | |
| | LEVEL-3 | INC10,HO2,INC23 | | | LEVEL-3 | LE13 | | |
| | | | | | | | | |
| OUTMOVE | LEVEL-0 | LE39 | 0.497 | 0.8521 | LEVEL-0 | LE39 | 0.498 | 0.8513 |
| | LEVEL-1 | INC6 | | | LEVEL-1 | INC6 | | |
| | LEVEL-2 | | | | LEVEL-2 | | | |
| | LEVEL-3 | | | | LEVEL-3 | | | |
| | | | | | | | | |
| UPMOVE | LEVEL-0 | LE21 | 0.571 | 0.7336 | LEVEL-0 | LE20 | 0.551 | 0.7339 |
| | LEVEL-1 | LE28 | | | LEVEL-1 | LE40 | | |
| | LEVEL-2 | LE16,INC19 | | | LEVEL-2 | LE28,INC23 | | |
| | LEVEL-3 | INC7,INC8,INC18 | | | LEVEL-3 | INC8,INC19 | | |

*Table 4.4.1.6. Phase2 common attributes with and without imputation*

## 4.5 Interpretation and Evaluation

After analyzing the various classification logs from Weka with our experiments on TGP dataset for waves 1 and 2, we have arrived at the following results.

***Dominant attributes identified for Wave 1:***

***Mig***: Life Events (LE28/GV16X28, LE25/GV16X25, LE16/GV16X16) and Income (INC18/BV32X14, INC22/BV34B)

***NoMig***: Area, Housing (HO8/ CV15X3), Income (INC22/BV34B, INC4/BV31, INC8/BV32X4), Life Events (LE28/GV16X28) and Attitudes towards relocation (AR6/FV8).

***AV1***: Attitudes towards relocation (AR6/FV8), Area, Income (BV28X8, INC8/BV32X4), Housing (HO8/CV15X3), Life Events (LE5/GV16X5), Health and Life Style (CV6X8), Social Connections (DV10X2)

***Dominant attributes identified for Wave 2:***

***Mig***: Attitudes towards relocation (AR5/F8, AR3/F6, AR4/F7)

***NoMig***: Attitudes towards relocation (AR5/F8, AR3/F6), Area

Looking at the above attributes we can observe that Life events; having lost someone in the family or relatives, Income of the person seemed to be dominant attributes that helps us predict the migration when mig(migrant) is used as decision label for wave 1. Similarly we have observed Area, housing availability, income, life events and attitudes towards relocation were identified as the dominant attributes for nomig(non-migrant) as decision label.

When a supplementary variable to mig(migrant) that is available in wave 1; i.e AV1 is used as decision variable the observed attributes were attitudes towards relocation, area, income, housing, life events, health & life style, and social connections. When we observed the patterns for mig and nonmig in wave 2, we have observed that only attitudes and area codes suffice to determine the labels. These results were identified to be valid and observed to meet the domain expert's expectations.

## 5. Observations and Insights

### 5.1 Observations

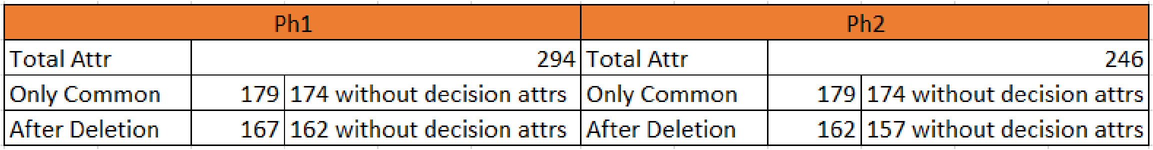
During the course of the project we have come across a few interesting observations.

i. The imputed data with the help of domain expert's suggestions with regards to rules for updating the data has very negligible effect on the classification rates of the samples in both wave 1 and wave 2.

ii. The classifier models developed in wave 2 have higher accuracy when we analyzed the data set using weka. When we reached out to the domain expert's evaluation on the same, it was found to be valid as the attributes recorded as responses to questionnaire after the relocation (wave 2) has depicted the definite and serious impacts on the lives of the people which are realistic.

## 5.2 Insights

To gain further insights into the dataset we have observed the common attributes through the questionnaire used for the survey. After identifying the common attributes, we have renamed the attributes in wave 1 & 2 to a common naming convention. We maintained the consistency of our analysis by only renaming the attributes that have recorded responses from both wave 1 and wave 2 for the same question. For instance, if the attribute in the data set belongs to attribute group: physical health (have you been diagnosed the past few days with any sort of illness), we have renamed it in both the phases to PH1. We have dealt with the variance in the number of attributes in both the phases by removing them, thus the dataset after the updates has reduced in just the number of attributes but not samples.



| Ph1 | | | Ph2 | | |
|---|---|---|---|---|---|
| Total Attr | 294 | | Total Attr | 246 | |
| Only Common | 179 | 174 without decision attrs | Only Common | 179 | 174 without decision attrs |
| After Deletion | 167 | 162 without decision attrs | After Deletion | 162 | 157 without decision attrs |

Table 5.2.1: Data statistics after identifying common attributes and renaming

Identifying common attributes in both phase 1 and phase 2 also did not lead to many differences in the dominant attributes being identified.

## 5.3 Threshold

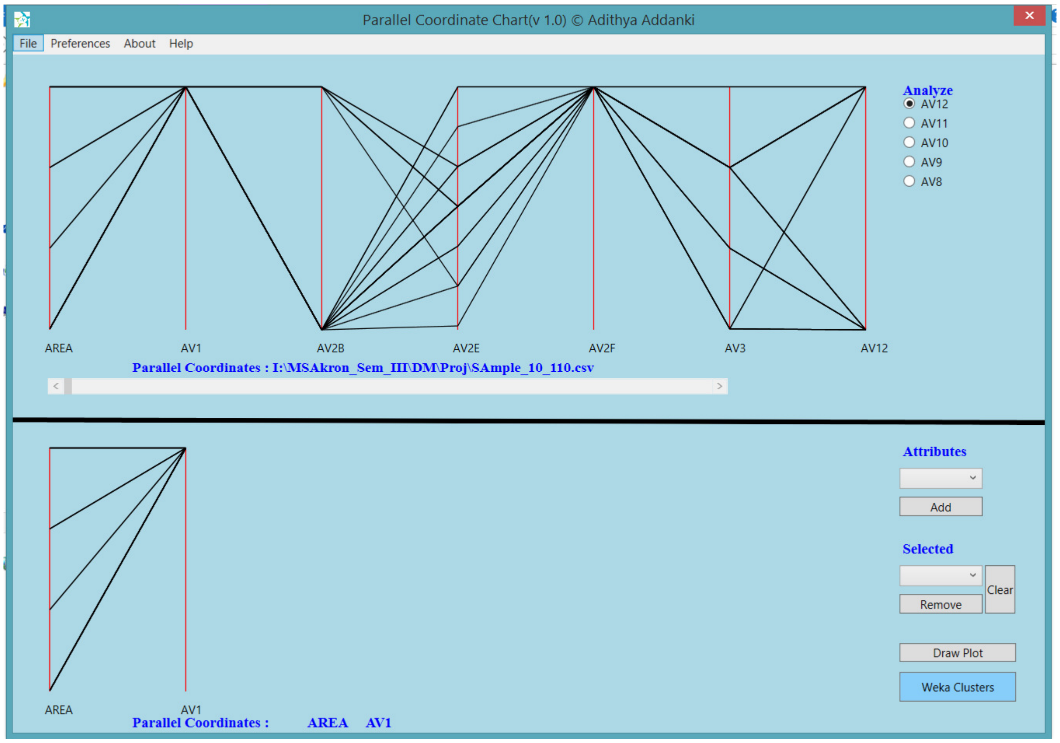
We have used 30% as the minimum threshold for missing values for all the attributes as per the domain expert's suggestion. We might have looked into the variations when the threshold is varied, possibly in the range of 10-50%.

# 6. Visual Analytics

## 6.1 Parallel Coordinates

Parallel coordinates are a very versatile and useful technique for finding structures in moderately-sized datasets. With a bit of experience, it is possible to very quickly recognize patterns and even estimate the strength of correlations, etc.

The implementation is done in WPF, with code behind programmed in C#.NET. The application has been designed and implemented to handle three kinds of files for parallel coordinate plots; CSV, XLS and XLSX. Also, once the plot is drawn the attributes could be filtered through user selection; brushing for selecting region of interest on the attributes. The axes could be reordered for gaining in-depth understanding on the distribution of data.



● AV12

Parallel Coordinates : I:\MSAkron_Sem_III\DM\Proj\SAmple_10_110.csv

Parallel Coordinates : AREA AV1

*Figure 6.1.1: Sample Dataset Loaded into Application*

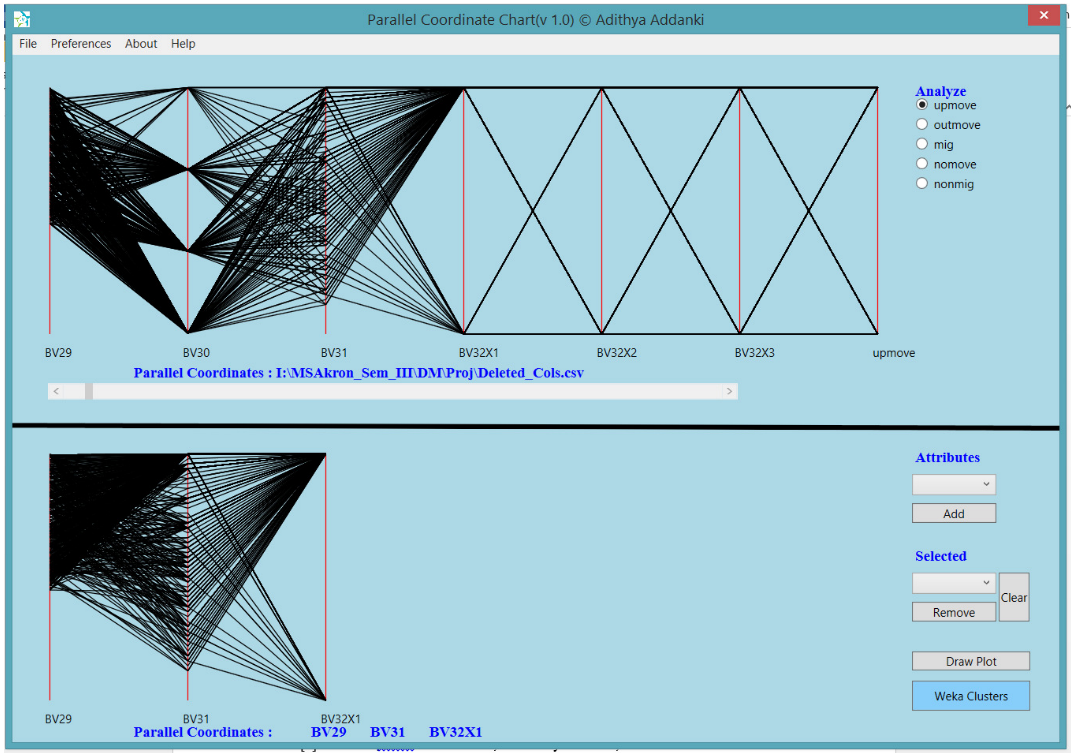


*Figure 6.1.2: TGP Dataset Loaded into Application*

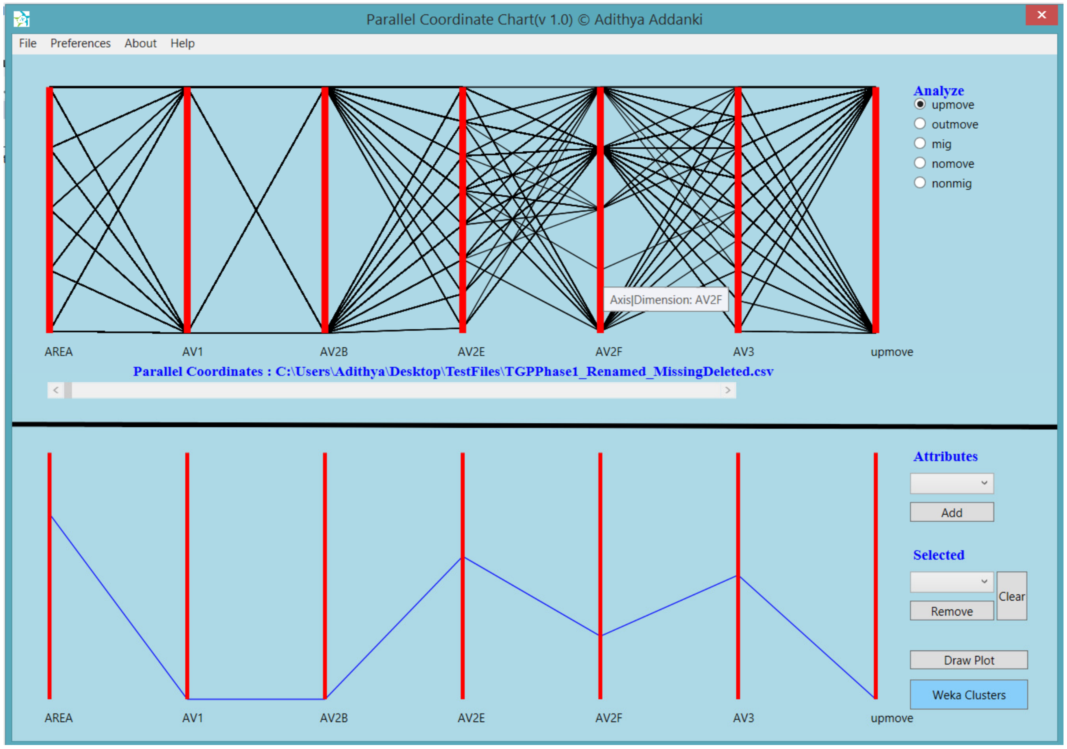


*Figure 6.1.3: Brushed plot: selected region of interest for Area.*

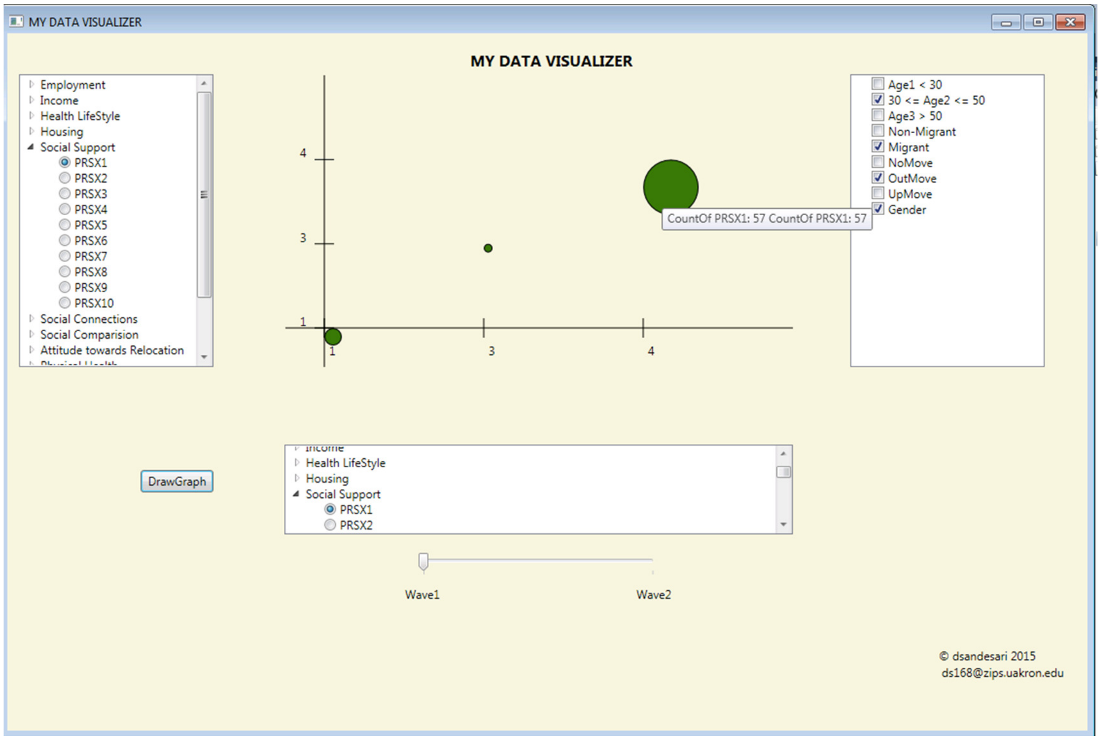
Representing heterogeneous data types and scaling the range of values to fit to fixed axes was a technical glitch. A few of limitations are listed below:

- Each axis can only be compared with two other axes at once;
  - One to the left another to the right
- More range of values in each attribute might cause confusion to the user.
  - Use with caution with regards to categorical data.
- ASCII based text characters only[character sets and digits]
- Rearrangement of attributes | axes order might be required to view patterns
  Good arrangement of axes could be dealt using heuristics.

### 6.2 Interactive Visualizer using Bubble Chart

There is a need of visualizer to interactively analyze the data. So an application is developed using WPF (XAML), C# as front-end and Microsoft SQL Server as backend that displays data from the database using bubble charts. The application consists of X-axis radio button selection list, Y-axis radio button selection list and checkbox decision list of attributes. It also consists of a ticker that helps in selecting the wave of database and a button that draws the graph. The data set Wave1 and Wave2 each consists of 183 attributes that are common to both phases including the decision attributes.

To analyze the data, for example select PRSX1 attribute pertaining to Social connection on both X-axis and Y-axis. The survey questionnaire for PRSX1 is that the respondents were asked, "Whether or not you could get help or assistance in the following areas on a regular basis if you needed it: 1) someone to lend you money to pay bills or help you get along?" Responses are coded 1 for "No", 2 for "Don't know", 3 for "Yes, with difficulty", and 4 for "Yes". To analyze a partition of data we can choose migrant females who out moved and aged between 30-50. From Figure. 6.2.1 and Figure 6.2.2. We can see analyze that the respondents who gave response 4, has been decreased from 57 in wave1 to 22 in wave2. So we can conclude that selected respondents received less help (about 50%) in wave2 that in wave1 i.e. the migration (due to TGP) had an effect on those respondents and affected their social connections.

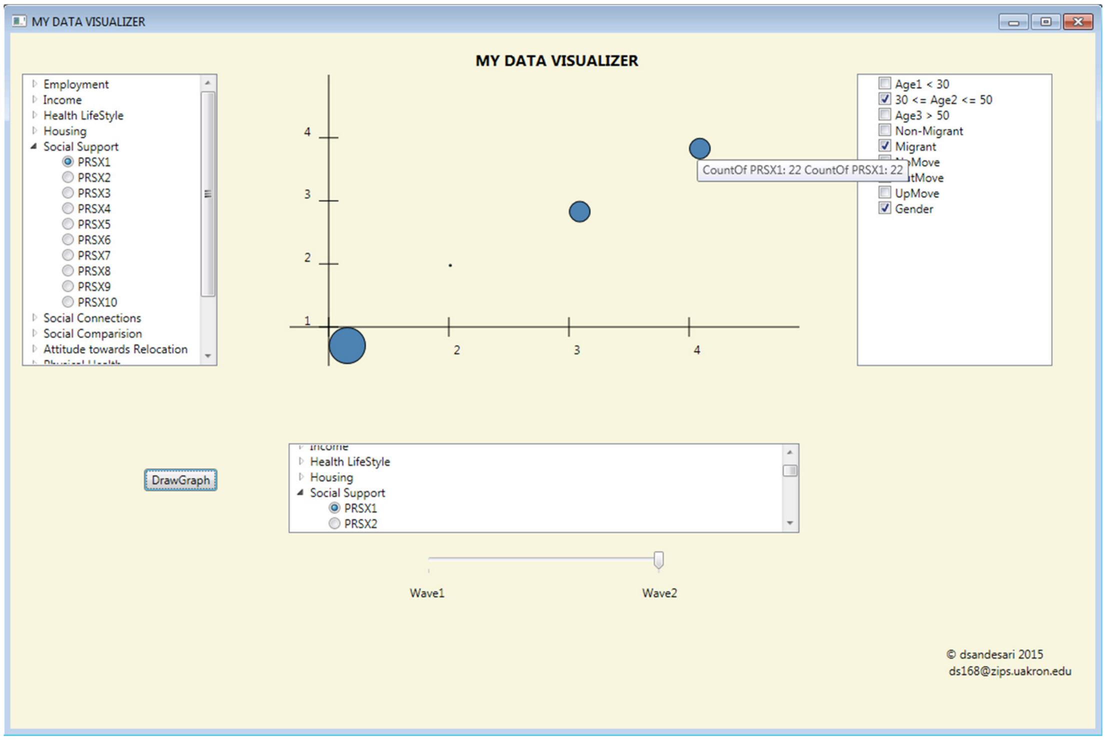


● PRSX1

☐ Age1 < 30
☑ 30 <= Age2 <= 50
☐ Age3 > 50
☐ Non-Migrant
☑ Migrant
☐ NoMove
☑ OutMove
☐ UpMove
☑ Gender

CountOf PRSX1: 57 CountOf PRSX1: 57

*Figure 6.2.1. Bubble plot visualization shown using data from wave1*



● PRSX1

☐ Age1 < 30
☑ 30 <= Age2 <= 50
☐ Age3 > 50
☐ Non-Migrant
☑ Migrant
☐ NoMove
☐ OutMove
☐ UpMove
☑ Gender

CountOf PRSX1: 22 CountOf PRSX1: 22

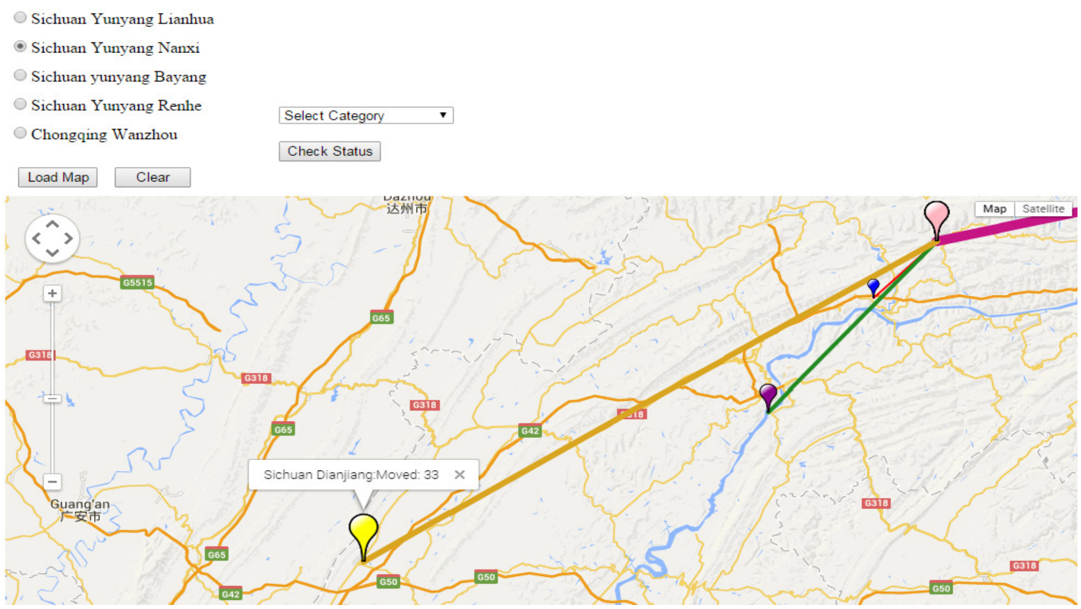
*Figure 6.2.2 Bubble plot visualization shown using data from wave2*

### 6.3 Geo-Visualization using Google Maps API

Geo-visualization is the geospatial analysis of data by using interactive visualization [15]. Geo-Visualization usually deals with creating static or dynamic interactive maps. It helps in data exploration and decision making processes. Initially Google provided merely JavaScript API for implementing maps, it is now extended to Adobe Flash and .NET applications. This application uses Google Maps API for geographic visualization, where it is potential to embed Google Maps in the website or webpage applications.

The main objective of this web-application is to geographically visualize the number of people forced to migrate from pre-migration location and to also visualize the effect of forced relocation from that particular place. The effects can be described as Perceived Routine Support, Self Esteem, Social Comparison, Mental Health and Physical Health statuses of people who were forced to migrate to different locations. This application is implemented using Asp.NET as Application Server, C# as Code behind Microsoft SQL Server 2012 as backend database, GMaps.dll and GoogleMaps.Subgurim.NET.dll files for embedding the Google Maps into application.

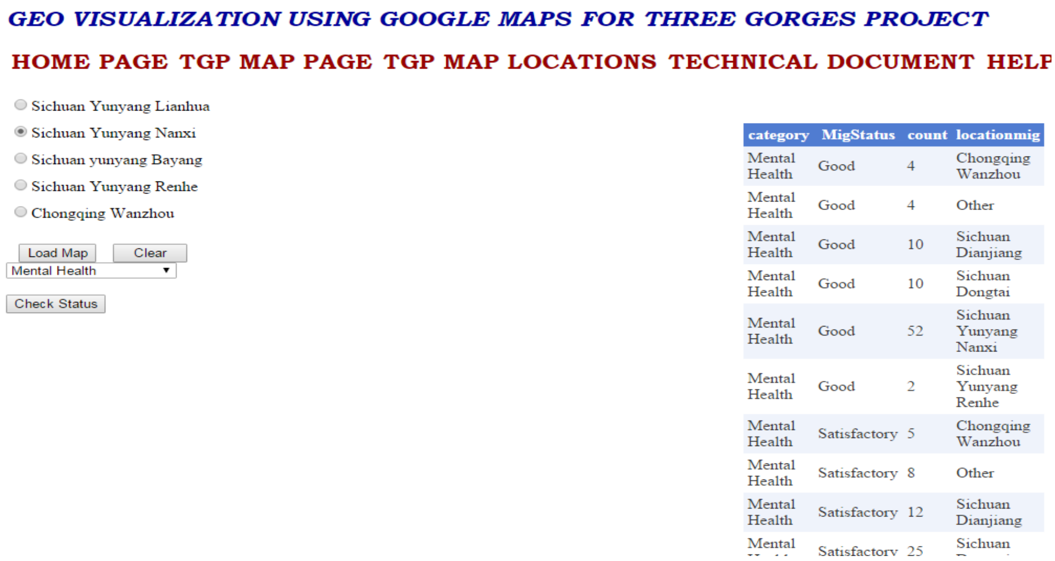
In this application, as shown in Figure 6.3.1. When a particular place is being selected from the pre-migration locations and LoadMap Button is clicked, the Google Maps is displayed with post-migration locations of people from that particular pre-migration location. The Marker size and marker color indicates number of people moved to that particular place. Even the polyline color and thickness changes based on the total number of people migrated from pre-migration location. Smaller the number of people migrated smaller is the size of marker and thin is the polyline weight.



Sichuan Dianjiang:Moved: 33 ✕

*Figure 6.3.1: Map displaying the post-Migration Locations based on number of people moved, marker size and color, polyline weight and color varies.*

Another feature of this application is shown in Figure 6.3.2, once the Map is displayed the user can select the category to which the status of migrated people can be shown. Upon selecting the category, web-application displays the status of people migrated to different locations.


GEO VISUALIZATION USING GOOGLE MAPS FOR THREE GORGES PROJECT
HOME PAGE TGP MAP PAGE TGP MAP LOCATIONS TECHNICAL DOCUMENT HELP
○ Sichuan Yunyang Lianhua
● Sichuan Yunyang Nanxi
○ Sichuan yunyang Bayang
○ Sichuan Yunyang Renhe
○ Chongqing Wanzhou
Load Map
Clear
Mental Health ▾
Check Status

| category | MigStatus | count | locationmig |
|---|---|---|---|
| Mental Health | Good | 4 | Chongqing Wanzhou |
| Mental Health | Good | 4 | Other |
| Mental Health | Good | 10 | Sichuan Dianjiang |
| Mental Health | Good | 10 | Sichuan Dongtai |
| Mental Health | Good | 52 | Sichuan Yunyang Nanxi |
| Mental Health | Good | 2 | Sichuan Yunyang Renhe |
| Mental Health | Satisfactory | 5 | Chongqing Wanzhou |
| Mental Health | Satisfactory | 8 | Other |
| Mental Health | Satisfactory | 12 | Sichuan Dianjiang |
| Mental | Satisfactory | 25 | Sichuan |

*Figure6.3.2: Web-application displays the Mental Health status of people migrated from Sichuan Yunyang Nanxi to different locations*

## 7. Future Scope and Conclusion

- By analyzing the dataset using weka through KDD process and visualization tools, we can identify the dominant attributes which predict the status of the migrants and non-migrants pre and post relocations.
- Data mining and visualization techniques could be extended as a template to other domain related datasets to identify hidden patterns in data and predict the characteristics/ features of samples.
- The visualizer proto-types developed for the project could be extended to full-fledged models so as to suit the needs of the domains.

## 8. References

[1] . Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1. http://www.cs.waikato.ac.nz/ml/weka/

[2] . http://www.cs.waikato.ac.nz/ml/weka/documentation.html

[3] . http://www.microsoft.com/en-us/server-cloud/products/sql-server/ - official site for Microsoft SQL Server.

[4] . Agrawala, M., Wilmot, L. I., & Berthouzoz, F. (2011). Design principles for visual communication. Communications of the ACM, 54(4), 60-69. doi:10.1145/1924421.1924439

[5] . http://en.wikipedia.org/wiki/Parallel_coordinates

[6] . http://www.xdat.org/ - open source tool in Java for Parallel Coordinates.

[7] . https://eagereyes.org/techniques/parallel-coordinates

[8] . Jain R.K., Kasana R.S., Jain S. (2009), Visualization of mined pattern and its human aspects, *International Journal of Computer Science and Information Security*, Vol. 4, No. 1&2, pp. 48–54.

[9] . D. A. Keim, C. Panse and M. Sips, Information Visualization: Scope, Techniques and Opportunities for Geovisualization. InJ. Dykes, A. MacEachren and M.-J. Kraak editors,Exploring Geovisualization. Oxford: Elsevier, 2004.

[10] . Google Charts, “Visualization: Bubble Chart”, https://developers.google.com/chart/interactive/docs/gallery/bubblechart

[11] . https://developers.google.com/maps

[12] . http://en.googlemaps.subgurim.net/

[13] . http://www.nullskull.com/q/10353401/show-address-on-google-map.aspx

[14] . http://www.codeproject.com/Articles/20590/Google-Maps-in-HTML-ASP-NET-PHP-JSPetc-with-ease

[15] . Lili Jiang, Qingwen Qi, & An Zhang. (2008). Study on GIS Visualization on Internet. Paper presented at the Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on, 1-4. doi:10.1109/WiCom.2008.2879

## 9. Statement of Work

***Addanki Adithya (aa207) [Team 1]:***

- ✓ Identified Phase 2 attributes and split the original set with Sharnitha.
- ✓ Acquired part of questionnaire (survey) from Dr. Juan Xi.
- ✓ Analyzed the missing values in phase2 and updated with '?'.
- ✓ Imputed phase 2 missing values with suggestions from Domain Expert.
- ✓ Identified common attributes in wave1 and wave2 with assistance from team and created classifier logs and statistics from Weka for common attributes in phase 1 and phase2.
- ✓ Documented results from Weka for analysis on common attributes in both the phases.
- ✓ Implemented Visualizer for parallel coordinate plot in WPF.

***Deekshith Sandesari (ds168) [Team 10]:***

- ✓ Identified Phase 1 attributes and split the original set into phase1 with Srinivas.
- ✓ Acquired part of questionnaire (survey) from Dr. Juan Xi.
- ✓ Imputed phase 2 missing values with suggestions from Domain Expert.
- ✓ Analyzed the missing values in phase1 and updated with '?'.
- ✓ Identified the common questionnaire in phase 1.
- ✓ Created classifier logs and statistics from Weka for phase 1.
- ✓ Documented results from Weka for analysis on phase1.
- ✓ Implemented visualizer for Bubble charts in WPF.

***Lakshmi Sharnitha Yarlagadda (ly27) [Team 10]:***

- ✓ Identified Phase 2 attributes and split the original set into phase2.
- ✓ Acquired part of questionnaire (survey) from Dr. Juan Xi.
- ✓ Analyzed the missing values in phase1 and updated with '?'.
- ✓ Imputed phase 1 missing values with suggestions from Domain Expert.
- ✓ Identified common attributes in phase1 and phase2 with assistance from team.
- ✓ Created classifier logs and statistics from Weka for phase2.
- ✓ Documented results from Weka for analysis on phase2.
- ✓ Implemented visualizer for Geo Visualization using Google Maps API and C#.

***Srinivasa Rao Katta (sk189) [Team 1]:***

- ✓ Identified Phase 2 attributes and split the original set into phase2.
- ✓ Acquired part of questionnaire (survey) from Dr. Juan Xi.
- ✓ Analyzed the missing values in phase2 and updated with '?'.
- ✓ Developed PL/SQL script for replacing the missing values with '?'.
- ✓ Imputed phase 1 missing values with suggestions from Domain Expert.
- ✓ Documented Imputation statistics for cases under analyses.
- ✓ Created distribution and imputation statistics from logs.