Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method

Apurwa Singh

2020

Thesis for the Degree of Master of Science in Computer Science

Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method

Apurwa Singh
2017-1-23-0014

Nepal College of Information Technology
Faculty of Science and Technology
Pokhara University, Nepal

January, 2021

Thesis for the Degree of Master of Science in Computer Science

Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method

Supervised by Roshan Koju, Ph.D.

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Computer Science

Apurwa Singh
2017-1-23-0014

Nepal College of Information Technology
Faculty of Science and Technology
Pokhara University, Nepal

January, 2021

DEDICATION

I would like to dedicate this thesis to my late grandparents who had more faith in me than I had in myself.

DECLARATION

I hereby declare that this study entitled **Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method** is based on my original research work. Related works on the topic by other researchers have been duly acknowledged. I owe all the liabilities relating to the accuracy and authenticity of the data and any other information included hereunder.

Apurwa Singh

2017-1-23-0014
Jan 12, 2021

RECOMMENDATION

This is to certify that this thesis entitled **Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method** prepared and submitted by **Apurwa Singh**, in partial fulfillment of the requirements of the degree of Master of Science (M.Sc.) in Computer Science awarded by Pokhara University, has been completed under my/our supervision. I/we recommend the same for acceptance by Pokhara University.

Dr. Roshan Koju
Social Security Fund
Jan 12, 2021

CERTIFICATE

This thesis entitled **Healthcare Vulnerability Mapping Using K-means++ Algorithm and Entropy Method** prepared and submitted by **Apurwa Singh** has been examined by us and is accepted for the award of the degree of Master of Science (M.Sc.) in Computer Science by Pokhara University.

Prof. Dr. Purushottam Kharel
Department of Computer Science
Nepal College of Information Technology
External examiner

Assoc. Prof. Dr. Roshan Chitrakar
Department of Computer Science
Nepal College of Information Technology
External examiner

Dr. Roshan Koju
Social Security Fund
Supervisor

Er. Niranjan Khakurel
Principal
Nepal College of Information Technology

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Dr. Roshan Koju for providing guidance and supervision during this research. Besides that, I would like to sincerely thank my dear friend Dr. Abhishek Dutta, Senior Research Software Engineer in the Visual Geometry Group (VGG) of the Department of Engineering Science at Oxford University. His support has been vital for this research. Likewise, I would also like to thank Arogya Koirala, Technical and Innovation Lead at Kathmandu Living Labs for his input in the initial phase of the research. I am also grateful to Ramu Khadkathoki, IT Officer of Ratnanagar Municipality for helping me get access to the source of data for this research. I would also like to thank Narayan Ban, Chief Administrative Officer of Ratnanagar Municipality for granting me access to the data for this research. Finally, I would like to thank Pravin Poudel for helping me with GIS maps.

Apurwa Singh
2017-1-23-0014
Jan 12, 2021

ABSTRACT

Vulnerable population in healthcare refer to those who are at greater risk of suffering from health hazards due to various socio-economic factors, geographical barriers and medical conditions. Mapping of this vulnerable population is a vital part of healthcare planning for any region. Results of vulnerability mapping later can help with meaningful interventions for healthcare demands.

This study focuses on combining geo analytics, unsupervised machine learning algorithm and entropy method for performing vulnerability mapping based on various above-mentioned factors.

In this study, k-means++ clustering algorithm is applied to household data of Ratnanagar municipality. Out of the available data, specific vulnerability indicators related to income, age, illness, disability and geolocation are created for the purpose of creating multiple clusters of households. One of the features used is, distance to the nearest health service provider, which is computed by using a routing engine called Open Source Routing Machine (OSRM), based on geolocations of households and health service providers of Ratnanagar. OpenStreetMap route data is used for this purpose.

After the clusters are formed, entropy method is used to evaluate vulnerability measure of each cluster. Later, based on population of different clusters in each ward and their respective vulnerability measures, each ward's vulnerability measure is quantified. Finally, a ward level vulnerability map is created for the municipality.

The results of this research can help decision makers perform evidence-based healthcare planning. The decision makers can identify the most vulnerable areas of the municipality and take rationale based mitigative measures.

Keywords: healthcare, vulnerability mapping, k-means++ clustering, elbow method, entropy method, OpenStreetMap, open source routing machine

Table of content

CHAPTER 5

List of Tables

List of Figures

List of Appendices

List of Abbreviations/Acronyms

CSV	Comma Separated Values
EM	Expectation Maximization
GIS	Geographic Information System
HTTP	Hypertext Transfer Protocol
IT	Information Technology
MOFAGA	Ministry of Federal Affairs and General Administration
OSRM	OpenStreet Routing Machine
OSM	OpenStreetMap
PCA	Principal Component Analysis
SoVI	Social Vulnerability Index
WCSS	Within Cluster Sum of Squares
XLS	Excel Spreadsheet

CHAPTER 1

INTRODUCTION

1.1 Background

Vulnerability is a state of being exposed to any potential harm. Vulnerability in healthcare is a measure of damage due to potential health issues. Various factors contribute to this vulnerability. These factors could be divided into following three categories:

i. Social-economic factors such as per capita income and age
ii. Geographic factor such as distance to the nearest health service provider
iii. Medical conditions such illness and disability

Some of the factors have a negative relationship with vulnerability while some have positive relationship. Out of the above-mentioned factors, per capita income has a negative relationship whereas the rest of the factors have a positive relationship with vulnerability.

Disability and illness also contribute to vulnerability. Due to disabilities, people have issues getting access to proper healthcare. Their disabilities make the interaction with the healthcare system even more difficult. Likewise, people with illness are already at risk of further deterioration of health.

Alongside the above-mentioned factors, age is also an important indicator of vulnerability in healthcare. Children and elderly are considered more vulnerable. Children are considered vulnerable because of their inability to protect themselves from any potential harm. Likewise, senior citizens are also considered vulnerable because of decreased immune system and physical capacity.

In line with above arguments, [1,2] mention that vulnerable population in healthcare include children, elderly, ill, disabled and socioeconomically underprivileged.

Healthcare vulnerability mapping is the process of creating a map which highlights parts of an area and their respective healthcare vulnerabilities.

In this study, ward is considered as the basic spatial unit. In order to perform vulnerability mapping, vulnerability indicators are prepared based on following parameters:

i. Family members' age
ii. Family members' illness status
iii. Family members' disability status

iv. Income

v. Geolocation

The indicators are prepared on a per household basis. After the indicators are prepared, clustering algorithm is applied to the data to create clusters of households based on similarity measure in terms of the vulnerability indicators.

For the purpose of clustering households, k-means++ clustering algorithm is used. The motivation behind choosing this particular algorithm is that k-means performs better with large data sets. [3] concludes that k-means algorithm is recommended for huge data sets. It is an unsupervised machine learning algorithm which can be used to identify clusters of data based on measure of similarity. The algorithm aims to minimize within cluster variance which is based on Euclidean distance. K-means++ is a specific version of the k-means algorithm which is used to improve the seeding process for the conventional k-means algorithm. This assures improvement in clustering results.

The clusters of households are further analyzed by entropy method to evaluate vulnerability index of individual clusters. In entropy method, objective weights are assigned to the vulnerability indicators based on information theory. Clusters are represented by centroid value of their respective indicators.

After evaluating vulnerability indices of clusters based on objective weights, ward level vulnerability index is computed and a vulnerability map is plotted as a GIS map.

1.2 Research Question

What are the levels of healthcare vulnerabilities of wards of Ratnanagar municipality based on combined effect of age, income, illness, disability and geolocation?

1.3 Research Objective

To evaluate vulnerability indices of wards and create ward level healthcare vulnerability map of Ratnanagar municipality based on combined effect of age, income, illness, disability and geolocation.

1.4 Significance of The Study

In context of Nepal, this study will be significant in performing statistical and strategic analysis in healthcare demands. As Nepal has entered into new federal structure and elected officials are on board after a long time, it is high time that socio-economic data is used for short-term and long-term planning and decision making. Based on the results of this research, local governments will be able to perform rationale-based healthcare planning. Besides that, the results will also serve as a baseline data for future comparisons.

1.5 Limitations of The Study

Since, one of the attributes used in this analysis is distance to the nearest health service provider, route data is vital for this research. In case of remote hilly areas of Nepal, a substantial amount of route data is missing. Thus, this study for now, is only limited to urban areas of Nepal where route data is already available in public domain such as OpenStreetMap. Besides that, data used for this research is based on household survey conducted by the municipality. Hence, the choice of vulnerability indicators was limited by the data availability.

1.6 Definition of Terms

Household: This refers to a family household which is a group of individuals who are related to each other and share a single residence. In some cases, this may also refer to people who are not related to each other but still they share a single residence.

Health Service Provider: This refers to different kinds of healthcare institutions such as clinics, health posts and hospitals who are licensed to provide various healthcare service.

CHAPTER 2

LITERATURE REVIEW

Various studies have been done on vulnerability mapping previously. These studies have studied different kinds of vulnerabilities. Based on different kinds of vulnerabilities, different indicators and frameworks have been proposed. Along with that, various weighting and summarizing methods have also been used for creating vulnerability index. However, all of these studies aim to produce vulnerability maps which help to make rationale-based decision making.

Social vulnerability index (SoVI) has been one of the most used tools for vulnerability index construction. The tool developed by Cutter et al. [4] has been used by multiple studies for vulnerability mapping [5-7]. SoVI uses PCA for quantifying vulnerability based on selected indicators of vulnerability. Originally, the tool was used for measuring social vulnerability to environmental hazards in United States.

In context of Nepal, due to lack of data availability very few quantitative vulnerability assessments have been done.

In [5], social vulnerability to natural hazards in Nepal was analyzed using a modified social vulnerability index. Using PCA, 7 factors were created out of 39 original indicators. The factors maintained 63.02% of the total variance. Kaiser normalization and Varimax rotation methods were used for selecting factors. Using Kaiser criterion, factors with eigenvalues greater than 1 were only selected. Finally, factors were aggregated using equal weighting method to form vulnerability index.

In [6], social vulnerability to natural hazards in Brazil was studied using social vulnerability index. 45 indicators were selected from 58 variables after testing for multicollinearity using Pearson's R calculation. After that, factor analysis of PCA using Kaiser normalization and Varimax rotation methods were performed to reduce the 45 indicators to 10 factors which represented about 67% of variance in data. Indicators which had more than 0.5 correlation or less than -0.5 correlation with the factors were considered as the drivers for the respective factors. Based on the drivers, cardinality of the components was set. Finally, the factors were summed using equal weights to product vulnerability index.

In [7], social vulnerability to flood hazards in Zimbabwe was studied using the same concepts as above. In this study, 17 variables were selected first which were then reduced to 4 factors using PCA. Factors were then merged into vulnerability index using equal weight method.

In [8], vulnerability to climate change in rural municipalities of Bosnia and Herzegovina was assessed. In this study, 20 indicators have been prepared for quantitative assessment of vulnerability to climate change. The indicators have been grouped into three components: exposure, sensitivity and adaptive capacity.

The study used an integrated approach which combined both potential impact of a hazard and adaptive capacity of a system. This approach considered vulnerability having both exogeneous and socio-economic dimensions.

The paper compared two weighting methods of equal weights and principal component analysis. In equal weight method, arithmetic mean of indicators for each component was calculated first. Next, arithmetic mean of indicators for exposure and sensitivity were summed to form sub-index for potential impact. Likewise, arithmetic mean of indicators belonging to adaptive capacity component was calculated to form sub-index for adaptive capacity. Finally, these two sub-indices were summed to form vulnerability index. Using equal weight method indicated that each indicator had equal contribution to overall vulnerability.

In second approach, the study used principal component analysis to create 6 factors from 20 indicators explaining 71.1% of variance in the data. Using Kaiser criterion only those factors were selected whose eigenvalues were greater than 1. Factor loadings were used as weight for the indicators. Based on these weights, vulnerability index was calculated as a weighted sum.

The paper suggested that using equal weight method is arbitrary and the results might be misleading. Using equal weight method implied that all the indicators had equal contribution to the overall vulnerability which might not be the case. The paper also suggested that using PCA for assigning weights based on loading factors might not be a guaranteed method as the loading factors were based on correlation and correlation did not necessarily represent causal relationship between indicators and vulnerability.

In [9], vulnerability mapping was done in regards to chemical hazards in the industrialized city of Shanghai. In this paper, genetic k-means algorithm was used for clustering civilian population based on specific attributes which represented exposure, sensitivity and coping capacity. In this study, min-max normalization was used. After cluster centroids were found, information entropy analysis was done to assign weight to various attributes for finding the most vulnerable cluster of population.

In [3], following clustering algorithms were compared: K-means Algorithm, Hierarchical Clustering Algorithm, Self-Organizing Map Algorithm and Expectation Maximization Clustering algorithm. Comparison was done based on following factors: size of dataset, number of clusters, type of dataset and type of software used. The paper concluded that EM and k-means algorithm are recommended for huge dataset.

Regarding data standardization, in [10], standardization methods were compared in terms of their effect on k-means clustering algorithm. Among z-score, decimal scaling and min-max normalization, min-max normalization was observed to be the best in terms of error sum of squares. An infectious diseases dataset with 15 data objects and 8 attributes was used to compare the standardization methods. K-means clustering algorithm with min-max normalization was found to have the minimum error sum of squares. Smaller error sum of squares assured higher accuracy in clustering. However, compared to Z-score method, min-max normalization had higher number of points which were out of cluster formation.

In [11], various k-value selection methods were compared. The paper compared four k-value selection algorithms of elbow method, gap statistic, silhouette coefficient and canopy algorithm. Based on execution time, elbow method seemed to be the most superior. In the study, for a dataset containing 100 samples, the execution time for elbow method was found to be 1.830 seconds. Next was canopy algorithm with execution time of 2.2120 seconds. In case of large and complex datasets, canopy algorithm seemed superior for its higher fault tolerance and noise immunity. However, canopy algorithm depended upon pre-defined distance thresholds and thus had more complexity than the elbow method.

CHAPTER 3

METHODOLOGY

3.1 Study Area

In context of Nepal, Ministry of Federal Affairs and General Administration (MOFAGA) has prepared Rural / Urban profile preparation guideline, 2074. The document guides local governments of Nepal to prepare a local government digital profile for the purpose of evidence-based planning. Local governments are first required to collect data from all its households and institutions based on a door-to-door survey. One key aspect of this survey is to collect geo data so that various geo analysis can be done.

For this study, case of Ratnanagar municipality was chosen. Ratnanagar municipality lies in Chitwan district of Bagmati province of Nepal. As per 2011 census [12], its estimated population is 46,607. Its total area is 35.62 square kilometers. Currently, there are 16 wards in Ratnanagar municipality.

3.2 Data Used

For this study, household survey data was chosen which included details regarding 17,501 households. Some of the important household information available in this data are as follows:

i. Ward number
ii. Annual income
iii. Family size
iv. Age of individual family members
v. Disability status of individual family members
vi. Health status of individual family members
vii. Geolocation

The raw household data was provided as an Excel workbook. The workbook contained following worksheets:

i. respondent
ii. member

The first worksheet contained details regarding individual household along with detail of the respondent. The respondent was responsible for providing all the household information and individual family member details. There were 17501 instances and 311 attributes in this worksheet. Each instance in this worksheet represented a single household.

The second worksheet contained details regarding the rest of the family members of the household. There were 66055 instances and 95 attributes in this worksheet. Each instance in this worksheet represented a single family member of a particular household.

So, for a household with 5 family members, the first worksheet contained a single instance with household and respondent data whereas, the second worksheet contained 4 instances for the rest of the family members.

Besides household data, a CSV file containing geolocations of health service providers was also prepared for the research.

3.3 Vulnerability Indicators

For the purpose of vulnerability assessment, following attributes were defined as vulnerability indicators. These attributes were derived from above mentioned data of household survey conducted by the municipality and were related to healthcare vulnerability [1,2]. It should be noted that selection of vulnerability indicators was influenced by availability of data:

1. Per capita income
2. Disability ratio
3. Illness ratio
4. Vulnerable age ratio
5. Distance to the nearest health service provider

1. Per capita income

Per capita income is calculated as follows:

$$per\ capita\ income = \frac{annual\ income}{family\ size} \tag{1}$$

Households with low per capita income are more vulnerable as the impact of adverse health conditions is higher for these households. Due to low income, they already may be deprived of quality healthcare in the first place. It should be noted that per capita income has negative correlation with vulnerability.

2. Disability ratio

Disability ratio is calculated as follows:

$$disability\ ratio = \frac{number\ of\ family\ members\ with\ disability}{family\ size} \tag{2}$$

Healthcare vulnerability increases with increase in disability ratio. Households with a higher disability ratio are more vulnerable as disability makes it more difficult to provide adequate healthcare due to accessibility issues. People with disabilities need special attention regarding healthcare as the general approach is not effective in such cases.

3. Illness ratio

Illness ratio is calculated as follows:

$$illness\ ratio = \frac{number\ of\ family\ members\ with\ some\ form\ of\ illness}{family\ size} \tag{3}$$

People with illness are at higher risk of further aggravation in health. One form of illness can lead to other forms of illness as well. Hence, healthcare vulnerability also increases with increase in illness ratio.

4. Vulnerable age ratio

Vulnerable age ratio is calculated as follow:

$$vulnerable\ age\ ratio = \frac{number\ of\ vulnerable\ age\ group\ members}{family\ size} \tag{4}$$

Vulnerable age group refers to family members who are more vulnerable in terms of their age. Children of age equal to and under 5 years and, elderly of age equal to and above 60 years are considered vulnerable family members. Thus, healthcare vulnerability also increases with increase in vulnerable age ratio. The reason for choosing these age limits is that globally child mortality rate refers to the mortality of child under age of five [13]. Regarding elderly age limit, in Nepal, citizens above age of 60 are considered senior citizens [14].

It is more appropriate to calculate ratio rather than count for illness, disability and vulnerable age because in a household, an individual can be ill and disabled, or can be ill and elderly at the same time and likewise. In such case, ratio is more effective than count in comparing different households.

5. Distance to the nearest health service provider

In case of a healthcare emergency, distance to the nearest health service provider becomes vital. Resources required to reach the health service provider increases with the distance. Thus, this also contributes to healthcare vulnerability.

Distance to the nearest health service provider is calculated for each household based on geolocation of the household and health service provider. For this, a routing server known as OpenStreet Routing Machine, OSRM was used. OSRM uses route data of OpenStreetMap, a free and open source map tool.

OSM data for Nepal is provided by Geofabrik, a consulting and software development firm based in Germany.

In order to find the distance to the nearest health service provider, distance to each health service provider was calculated for a given household. After that, the smallest value was chosen as the distance to the nearest health service provider.

3.4 Proposed Model

As seen in the literature review section, vulnerability mapping studies [5,6,7] have used Social Vulnerability Index in modified forms. SoVI utilizes dimension reduction technique of Principal Component Analysis. PCA becomes useful when there are high number of variables to start with. Since, the number of variables used in this study is not high, using PCA is not appropriate and beneficial.

Studies have also suggested that using PCA for index construction might not be accurate because if the weighting mechanism is based on correlation it might not quantify the impact of sub indicators on the vulnerability accurately [15]. An alternative to using PCA would be to use machine learning algorithm of clustering followed by entropy method for weighting purpose.

The advantage of using clustering algorithm is that it maintains transparency of dimensions of vulnerability [9]. Vulnerability of two locations may be same in magnitude when quantified but they may have different composition in terms of indicators. Thus, using clustering algorithm makes sure that dimensions of the vulnerability are discernible which can help with effective vulnerability mitigation.

Results of clustering algorithm can be analyzed by entropy method during which objective weights are assigned to indicators based on their entropy.

In previous vulnerability studies, geo analytics was not significantly involved. In [9], one of the vulnerability indicators used was distance to the nearest main road. However, the paper doesn't give details about how the distance was computed.

Since the household survey included geolocations of all the households, it was an opportunity to involve geo analytics in the vulnerability assessment. Hence, using feature engineering, distance to the nearest health service provider was established as one of the indicators of healthcare vulnerability.

Thus, in this study, geo analytics, machine learning and information theory are combined in order to evaluate vulnerability index of wards of the municipality.

The model can be divided broadly into 3 parts: (i) data preprocessing and feature engineering, (ii) clustering and, (iii) entropy method and vulnerability index calculation. Detailed information about the model is explained below.

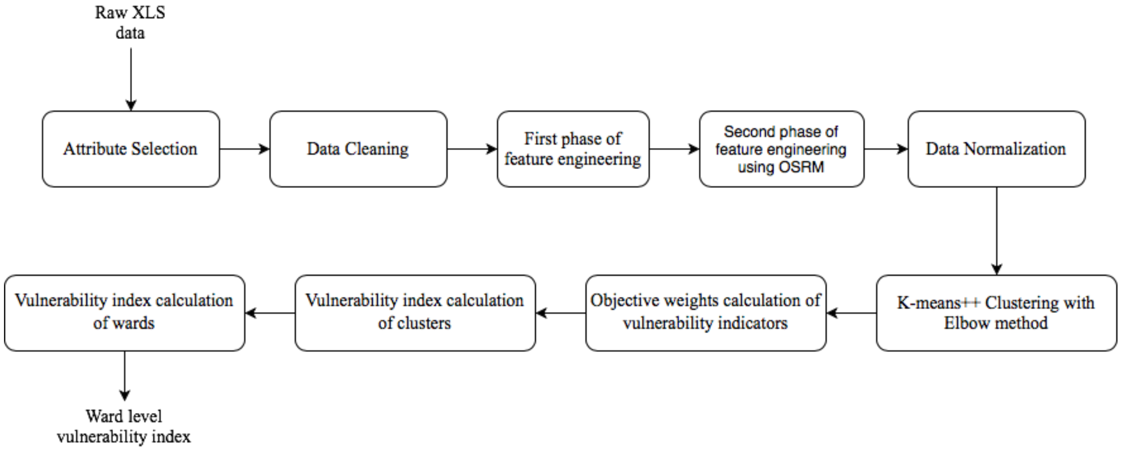

Raw XLS data
Attribute Selection
Data Cleaning
First phase of feature engineering
Second phase of feature engineering using OSRM
Data Normalization
K-means++ Clustering with Elbow method
Objective weights calculation of vulnerability indicators
Vulnerability index calculation of clusters
Vulnerability index calculation of wards
Ward level vulnerability index

Figure 1. Proposed model

The model used for this study is shown above. The components of the model are as follows:

i. Attribute Selection
ii. Data Cleaning

iii. First phase of feature engineering

iv. Second phase of feature engineering

v. Data Normalization

vi. K-means++ clustering with elbow method

vii. Objective weights calculation of vulnerability indicators

viii. Vulnerability index calculation of clusters

ix. Vulnerability index calculation of wards

Attribute selection, data cleaning and normalization are parts of data preprocessing.

3.4.1 Attribute Selection

In the respondent worksheet, following household and respondent attributes were selected:

Table 1. Household attributes

S. no	**Column**	**Attribute**	**Remarks**	**Data type**
1	A	ward	Ward location of household	Discrete numerical
2	B	family_size	Count of family members	Discrete numerical
3	G	annual_income	Annual income of household in NPR	Continuous numerical
4	H	latitude	Latitude part of household geolocation	Continuous numerical
5	I	longitude	Longitude part of household geolocation	Continuous numerical
6	J	_id	Unique identifier for individual household survey	Discrete numerical

Table 2. Respondent attributes

S.no	**Column**	**Attribute**	**Remarks**	**Data type**
7	C	r_age	Age of respondent in years	Discrete numerical
8	D	r_age_unit	Unit of respondent age	Nominal categorical
9	E	r_disability	Disability status of respondent	Categorical
10	F	r_ill	Health status of respondent	Categorical

Likewise, in the member worksheet following attributes were selected for individual family members:

Table 3. Family member attributes

S.no	Column	Attribute	Remarks	Data type
1	A	m_age	Age of family member in years	Discrete numerical
2	B	m_age_unit	Unit of family member age	Categorical
3	E	m_disability	Disability status of family member	Categorical
4	F	m_ill	Health status of family member	Categorical
5	J	_submission__id	Unique identifier for individual household survey	Discrete numerical

_id and _submission_id attribute values are used to associate rows of first worksheet with the rows of second worksheet for individual households.

By selecting only relevant attributes in the first place, it was easier to work with the large dataset throughout the research.

3.4.2 Data Cleaning

The raw data contained noisy data and some data were missing as well. By observing the values of attributes latitude and longitude, noisy data were identified. This was also verified by plotting the coordinates on a map. It was observed that the geolocations lied outside the study area.

11 instances of data were noisy and 1 instance had no data for the attributes. Thus, 12 instances of data were removed from the first worksheet.

Since, the number of such instances were very insignificant compared to the total instances in the worksheet, it was safe to remove these instances. Finally, the number of instances in the first worksheet were 17,489.

There were multiple missing data for the attribute annual_income. These values were replaced by arithmetic mean of known attribute values. The average value was found to be 354621.3504.

3.4.3 Feature Engineering

Feature engineering is the process of creating new features from existing ones for improving the results of machine learning algorithm. By using feature engineering, the input to the machine learning algorithm can be improved which leads to better results.

After performing attribute selection and data cleaning on the data set, new features were created from existing ones in two phases of feature engineering.

3.4.3.1 First phase of feature engineering

In the beginning of the first phase, following features were created in the member worksheet which contained details regarding family members excluding the respondent:

Table 4. Family member attributes after feature engineering

S.no.	Attribute	Detail
1	m_disability_numeric	Disability status of an individual family member excluding the respondent. It was obtained by converting categorical attribute m_disability to numeric form using integer encoding.
2	m_ill_numeric	Health status of an individual family member excluding the respondent. It was obtained by converting categorical attribute m_ill to numeric form using integer encoding.
3	m_vulnerable_age_numeric	Age related vulnerability status of an individual family member excluding the respondent. It was obtained by checking if age of the family member was less than or equal to 5 years or, more than or equal to 60 years.
4	m_disability_total	Total number of family members excluding the respondent who had some form of disability.
5	m_ill_total	Total number of family members excluding the respondent who had some form of illness.
6	m_vulnerable_age_total	Total number of family members excluding the respondent who are vulnerable in terms of their age.

Table 5. Excel formula for family member attributes

S.no.	Attribute	Excel Formula	Remarks
1	m_disability_numeric	IF(E2="अपाङ्गता नभएको",0,1)	E2 = m_disability
2	m_ill_numeric	IF(F2="स्वस्थ",0,1)	F2 = m_ill
3	m_vulnerable_age_numeric	IF(AND(A2>=1,A2<=5),1,IF(AND(A2<=60,B2="महिना"),1,IF(AND(A2>=60,B2="वर्ष"),1,0)))	A2 = m_age, B2 = m_age_unit
4	m_disability_total	SUMIF(J:J,J2,K:K)	J = _submission_id, K = m_disability_numeric
5	m_ill_total	SUMIF(J:J,J2,L:L)	J = _submission_id, L = m_ill_numeric
6	m_vulnerable_age_total	SUMIF(J:J,J2,M:M)	J = _submission_id, M = m_vulnerable_age_numeric

It should be noted that all the Excel formula in this research have been expressed for row number 2.

Later, the last three features mentioned above were obtained in the first worksheet. For this purpose, VLOOKUP functions were used in the formula as follows:

Table 6. Excel formula for family member attributes to be obtained in first worksheet

Attribute	Formula
m_disability_total	IF(B2=1,0,VLOOKUP(J2,member!J:N,5,FALSE))
m_ill_total	IF(B2=1,0,VLOOKUP(J2,member!J:O,6,FALSE))
m_vulnerable_age_total	IF(B2=1,0,VLOOKUP(J2,member!J:P,7,FALSE))

Here, VLOOKUP function was only used when the number of family members was more than 1. This was checked by IF function.

After computing attributes for rest of the family members, household and respondent features were created in the first worksheet as follows:

Table 7. Household and respondent attributes after feature engineering

S.no.	Attribute	Detail
1	per_capita_income	Annual income of an individual household per capita. It was calculated by combining two features annual_income and family_size.
2	r_disability_numeric	Disability status of respondent. It was obtained by converting categorical attribute r_disability to numeric form using integer encoding.
3	r_ill_numeric	Health status of respondent. It was obtained by converting categorical attribute r_ill to numeric form using integer encoding.
4	r_vulnerable_age_numeric	Age related vulnerability status of respondent. It was obtained by checking if age of respondent was more than or equal to 60 years. It was safe to assume that no respondent could be of age less than or equal to 5 years.
5	disability_total	Total number of family members in a household with some form of disability.
6	ill_total	Total number of family members in a household with some form of illness.
7	vulnerable_age_total	Total number of family members in a household who are vulnerable due to their age.
8	disability_ratio	Ratio of total number of family members with disability to family size.
9	ill_ratio	Ratio of total number of family members with illness to family size.
10	vulnerable_age_ratio	Ratio of total number of age wise vulnerable family members to family size.

Table 8. Excel formula for household and respondent attributes

S.no.	Attribute	Excel Formula	Remarks
1	per_capita_income	G2/B2	G2=annual_income, B2=family_size
2	r_disability_numeric	IF(E2="अपाङ्गता नभएको",0,1)	E2 = r_disability
3	r_ill_numeric	IF(F2="स्वस्थ",0,1)	F2 = r_ill
4	r_vulnerable_age_numeric	IF(C2>=60,1,0)	C2 = r_age
5	disability_total	SUM(M2,P2)	M2 = r_disability_numeric, P2 = m_disability_total

6	ill_total	SUM(N2,Q2)	N2 = r_ill_numeric, Q2 = m_ill_total
7	vulnerable_age_total	SUM(O2,R2)	O2 = r_vulnerable_age_numeric, Q2 = m_vulnerable_age_total
8	disability_ratio	S2/B2	S2=disability_total, B2 = family size
9	ill_ratio	T2/B2	T2 = ill_total, B2 = family size
10	vulnerable_age_ratio	U2/B2	U2 = vulnerable_age_total, B2 = family_size

Finally, a CSV file was created by selecting following household attributes:

i. ward
ii. latitude
iii. longitude
iv. per_capita_income
v. disability_ratio
vi. ill_ratio
vii. vulnerable_age_ratio

This marked the end of first phase of feature engineering.

3.4.3.2 Second phase of feature engineering

In the second phase of feature engineering, distance to the nearest health service provider was computed using OpenStreet routing machine.

In order to calculate distance to the nearest heath service provider, geolocation of households and health service providers had to be used along with the route data. For this purpose, Open Street Routing Machine was used. OSRM used road network data of OpenStreetMap data extracts. OSRM was installed as a docker container in local machine.

After doing all the necessary configurations, HTTP request could be made against the local running OSRM server to find out the shortest distance between two geolocations based on route data.

Let’s consider two geolocations A and B given by following GPS coordinates:

A: 27.6638477, 84.54705958

B: 27.6225471, 84.5187781

If one were to compute shortest path between the two points A and B, a sample HTTP request to OSRM server would be as follows:

http://127.0.0.1:5000/route/v1/driving/84.54705958,27.6638477;84.5187781,27.6225471?overview=false

In this case the OSRM server is running in local machine at port 5000.

The response to the request is as follows:

{"code":"Ok","routes":[{"legs":[{"steps":[],"**distance**":**6996.5**,"duration":920.8,"summary":"","weight":920.8}],"distance":6996.5,"duration":920.8,"weight_name":"routability","weight":920.8}],"waypoints":[{"hint":"idUFgIvVBYA0AAAAFAAAAAAAAADuAAAAZ_UQQhKNVkEAAAAAnC4lQzQAAAAUAAAAAAAAAO4AAAAnAAAAUBYKBecdpgH0FQoF6B2mAQAADwAITebR","distance":9.077145,"name":"","location":[84.547152,27.663847]},{"hint":"SKwFgE6sBYAZAAAAIQAAAAgBAABiAAAAOwKPQZ9Xs0H06DZDIr6HQhkAAAAhAAAACAEAAGIAAAAnAAAACqkJBc97pQF6pwkFk3ylAQcAzwEITebR","distance":45.060339,"name":"","location":[84.519178,27.622351]}]}

From the response, it can be observed that the shortest distance between the two locations is 6996.5 meters.

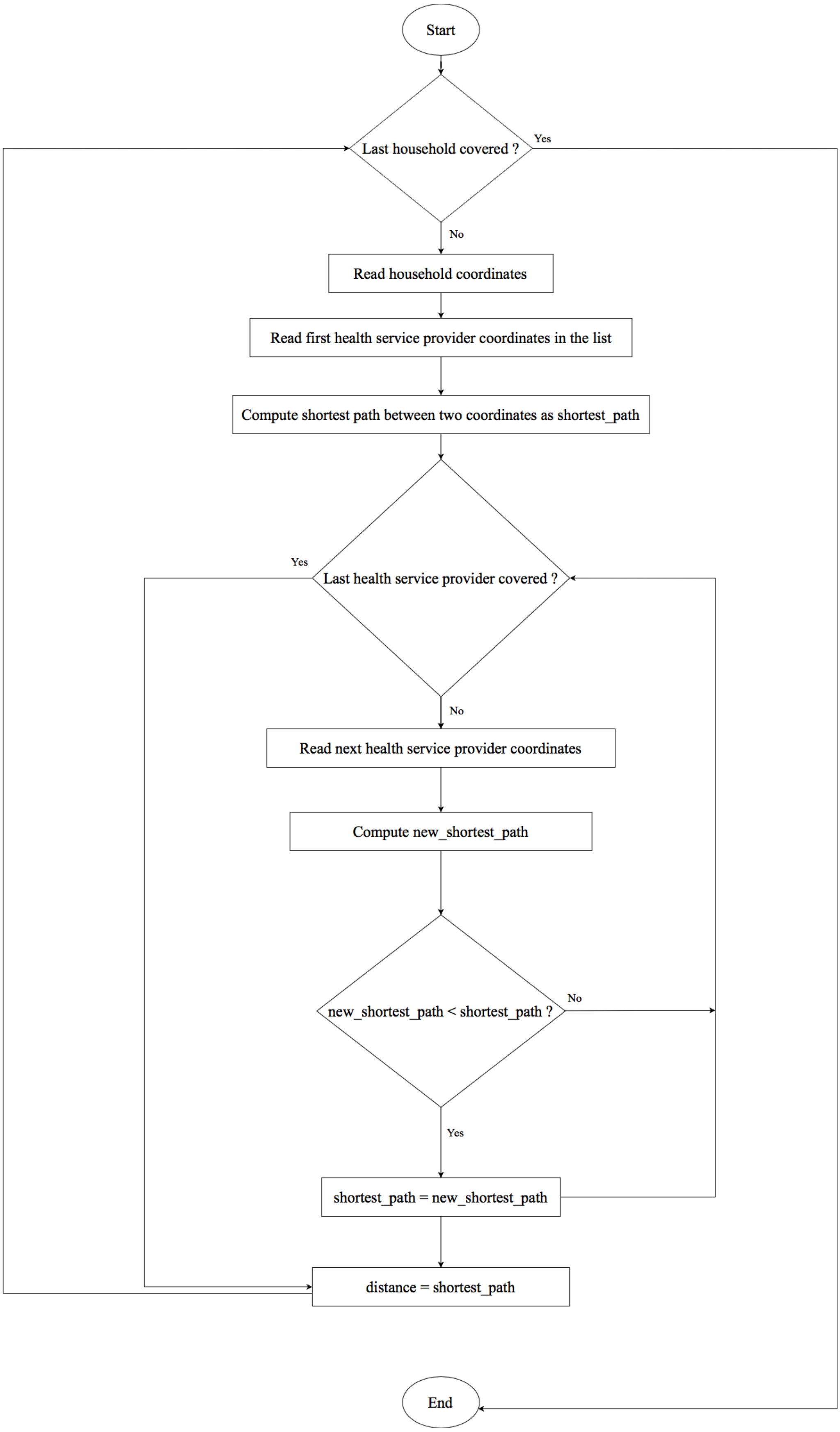

Start
Last household covered ?
Yes
No
Read household coordinates
Read first health service provider coordinates in the list
Compute shortest path between two coordinates as shortest_path
Yes
Last health service provider covered ?
No
Read next health service provider coordinates
Compute new_shortest_path
new_shortest_path < shortest_path ?
No
Yes
shortest_path = new_shortest_path
distance = shortest_path
End

Figure 2. Algorithm for finding distance to the nearest health service provider

3.4.4 Data Normalization

To make sure that no attributes became dominating in cluster analysis, normalization had to be applied to the dataset. In absence of normalization, a feature with larger scale contributes more to the cluster formation than features with smaller scale. In case of our data, it could be seen that per capita income had a larger scale compared to other features. Thus, it had to be made sure that data was normalized before the clustering algorithm took over.

Min-max normalization and z-score normalization are two of the most popular normalization methods.

In this study, min-max normalization was chosen as the standardization method, because min-max normalization assures that no values are transformed into negative values.

Entropy method is used later to find out the objective weights of the vulnerability indicators. Since the entropy method uses logarithmic functions to evaluate the objective weights, negative values cannot be used, as logarithm of a negative number in undefined. Z-score normalization transforms a value into a negative number if the number is smaller than the mean of the dataset. Hence, z-score normalization cannot be used with the entropy method.

In min-max normalization, values are transformed as follows:

$$Minmax(x) = \frac{\mathrm{x} - \min}{\max - \min} \tag{6}$$

where, min = minimum value in dataset and, max = maximum value in the dataset

Due to this, the maximum value in the dataset is transformed into 1 whereas the smallest value is transformed to 0. Thus, the whole dataset is transformed to values in the range (0,1).

In case of per capita income, the formula for min-max normalization was modified as follows as it had negative relationship with healthcare vulnerability:

$$Minmax(x) = 1 - \frac{\mathrm{x} - \min}{\max - \min} \tag{7}$$

3.4.5 Clustering

After performing data preprocessing and feature engineering, k-means++ clustering algorithm was applied to the data.

3.4.5.1 K-means and K-means++ algorithm

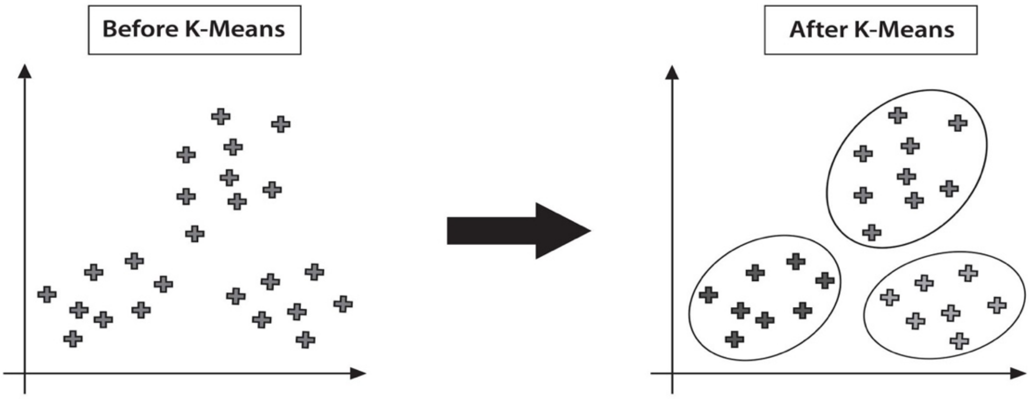


Figure 3. K-means clustering algorithm

Before we learn about k-means++ algorithm, k-means algorithm and its drawbacks are to be understood. K-means is an unsupervised machine learning algorithm used for cluster analysis. It is known for its simplicity, robustness and ability to work effectively with large data sets [3]. K-means is used to partition n samples into k sets such that each sample belongs to a specific cluster. The samples belonging to any given cluster are more similar to each other as compared to samples of any other clusters. In other words, k-means aims to reduce intra-cluster distance. Each cluster has its own centroid, which is the arithmetic mean of all the samples belonging to that cluster.

The algorithm works as follows:

1. Select k as number of clusters to be formed.
2. Randomly select k samples as initial centroids. These initial samples are also called seeds.
3. For each remaining sample, calculate the distance to all k centroids. Based on the distance, assign each sample to the nearest centroid.
4. For the new cluster formed, recompute the new centroid based on arithmetic mean of the samples.

5. Repeat steps 3 and 4 until the new centroids do not change significantly in value.

There are two drawbacks of naive k-means algorithm. They are:

1. The result of clustering depends upon the initial random selection of samples as centroids. Inappropriate initialization of centroids can lead to bad clustering results.
2. The result of clustering also depends upon value of K. Inappropriate value of K can also lead to bad results.

K-means++ algorithm is used to address the first limitation of the conventional k-means algorithm. Instead of choosing k random centroids, k-means++ only chooses one random centroid. Rest of the centroids are chosen on the basis of distance. After k centroids are finalized, one could proceed with the conventional k-means algorithm. The algorithm works as follows:

1. Randomly choose a sample as first centroid.
2. For each remaining sample, calculate the distance to the first centroid.
3. Select a sample as new centroid such that probability of selection is directly proportional to the square of the distance from the nearest centroid.
4. Repeat 2 and 3 until all the centroids have been chosen.
5. After K centroids are chosen, continue with the conventional k-means algorithm.

Thus, it can be seen that, k-means++ algorithm is only involved with centroids initialization. K-means++ algorithm makes sure that randomness in initialization is reduced which leads to less arbitrary results.

3.4.5.2 Elbow Method

To address the second issue of k-means, which is determination of optimum value of k, elbow method was used.

In elbow method, sum of squared distances of samples to the nearest cluster centroids, also known as within cluster sum of squares (wcss) or inertia, is plotted against the number of clusters.

$$wcss = \sum_{i=1}^{n} \sum_{x \in k_i} d(x, c_i)^2 \tag{8}$$

where, n=number of clusters, k_i = i^{th} cluster, c_i is the centroid of i^{th} cluster, x is a member of i^{th} cluster and d is a Euclidean distance function.

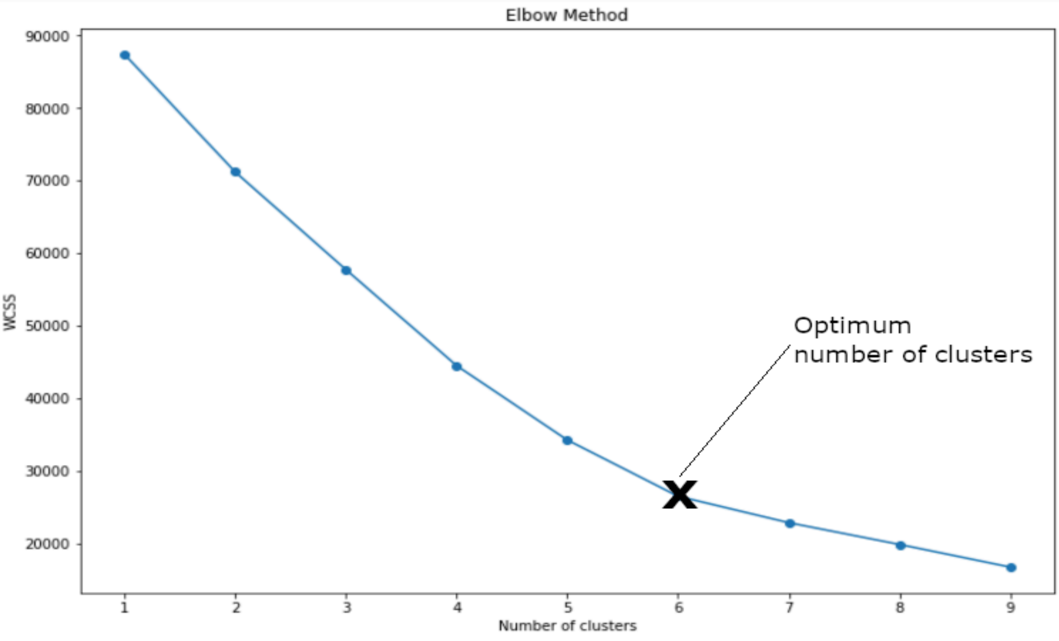


Figure 4. Elbow method

Initially, inertia decreases rapidly with increase in number of clusters. At a certain value of clusters, change in inertia becomes more stable and linear. This value of clusters was chosen as number of clusters to be formed from the algorithm.

3.4.6 Objective Weight Calculation of Vulnerability Indicators

Clustering leads to a set of clusters with different values for centroids of respective attributes. Now, in order to compare the clusters with each other, weightage have to be assigned to all the attributes.

Entropy method is used to assign objective weights to the attributes of clustering. Objective weights are solely based on inherent statistics rather than anyone’s judgement, preference or opinion.

In Shannon’s information theory [16], information entropy is the average value of unpredictability, uncertainty, surprise or information associated with a random variable.

Information is conveyed when an outcome of a random variable is identified. More improbable the outcome, more information is conveyed. In other words, one doesn't gain much information from knowing about occurrence of a highly probable event whereas, more information is gained when an unlikely event occurs. The quantified measure of this information or surprise associated with a single outcome is known as self-information. Information entropy is the expected value of self-information.

For a given data set, lower entropy leads to higher degree of diversification. For an attribute, degree of diversification is the measure of how disparate centroid values are in magnitude. It's easy to decide between values which are more disparate. Hence the attribute contributes more to the process of decision making. Likewise, when the attribute has high entropy, values are less disparate and decision making becomes more difficult due to smaller margin.

Due to this effect of degree of diversification on decision making, objective weight is calculated as normalized value of degree of diversification.

The steps followed in entropy method are as follows:

1. Create normalized decision matrix

A decision matrix is prepared where rows correspond to clusters and columns to vulnerability indicators. First, the decision matrix is normalized.

For an attribute j, normalized values are calculated as,

$$n_{ij} = \frac{x_{ij}}{\sum_{i=1}^{k} x_{ij}} \tag{9}$$

where, x_{ij} is the centroid value of i^th^ cluster and j^th^ vulnerability indicator and k is the number of clusters.

2. Calculate entropy of attributes

As per Shannon's Information theory [11], for attribute j, entropy is given by:

$$e_j = -\frac{1}{\ln(k)} \sum_{i=1}^{k} n_{ij}\, ln(n_{ij}) \tag{10}$$

where, k is the number of clusters and $0 \leq e_j \leq 1$

3. Calculate degree of diversification

It can be calculated from entropy as follows:

$$d_j = 1 - e_j \tag{11}$$

4. Calculate objective weight

For vulnerability indicator j, objective weight is calculated as normalized degree of diversification as follows:

$$w_j = \frac{d_j}{\sum_{j=1}^{k} d_j} \tag{12}$$

where, k is the number of clusters

3.4.7 Vulnerability Index of Clusters

After obtaining objective weights of the vulnerability indicators, vulnerability index of a cluster is calculated as follows:

$$v_i = \sum_{j=1}^{k} w_j * c_j \tag{13}$$

where, v_i is the vulnerability measure of cluster i, w_j and c_j are weight and centroid value for cluster i and vulnerability indicator j.

3.4.8 Vulnerability Index of Wards

Finally, vulnerability index of a ward can be calculated from vulnerability indices of clusters as follows:

$$x_i = \frac{\sum_{j=1}^{k} v_j * p_j}{p_i} \tag{14}$$

where, x_i is the vulnerability index of ward i, k is the number of clusters, v_j is the vulnerability index of cluster j, p_j is the number of cluster j households in ward i and p_i is the total number of households in ward i.

Thus, if a ward has x, y and z number of households in clusters 1,2 and 3 and, the clusters' vulnerability indices are w_1, w_2 and w_3 respectively, then vulnerability index of the ward is:

$$\frac{\mathrm{w}_1 x + \mathrm{w}_2 y + w_3 z}{x + y + z} \tag{15}$$

3.5 Source of Data

The household data of Ratnanagar municipality was made available on request to access the data for research purpose. A formal letter was sent to the municipality on April 21, 2020 to which the municipality replied on May 08, 2020 with another letter. Health service provider geolocations were collected from OpenStreetMap portal with verification from the municipality.

3.6 Tools

Following tools were used for this research:

1. Microsoft Excel

   Microsoft Excel was used to perform data preprocessing, first phase of feature engineering, objective weight calculation and vulnerability index calculation.
2. Anaconda with Jupyter notebook

   Jupyter notebook was used to write Python program for second phase feature engineering, normalization, clustering and further data analysis.
3. Docker

   Docker was used to run OpenStreet Routing Machine server.
4. OpenStreet Routing Machine

   OpenStreet Routing Machine was used to find the distance between geolocations in order to calculate distance to the nearest health service provider.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Results

Table 8 and 9 show datasets after first and second phase of feature engineering respectively.

Table 9. Dataset after first phase of feature engineering

	ward	**latitude**	**longitude**	**per_capita_income**	**disability_ratio**	**ill_ratio**	**vulnerable_age_ratio**
count	17489	17489	17489	17489	17489	17489	17489
mean	9.258277	27.628211	84.515090	8.035393e+04	0.005353	0.029109	0.173640
std	4.482088	0.024836	0.017609	1.551861e+05	0.040333	0.105522	0.215523
min	1.000000	27.573626	84.468262	3.333333e+02	0.000000	0.000000	0.000000
25%	5.000000	27.613020	84.502555	3.000000e+04	0.000000	0.000000	0.000000
50%	10.000000	27.629432	84.512678	5.000000e+04	0.000000	0.000000	0.142857
75%	13.000000	27.647457	84.527758	9.000000e+04	0.000000	0.000000	0.285714
max	16.000000	27.680680	84.557053	1.000000e+07	1.000000	1.000000	1.000000

Table 10. Dataset after second phase of feature engineering

	per_capita_income	**disability_ratio**	**ill_ratio**	**vulnerable_age_ratio**	**nearest_distance**
count	17489	17489	17489	17489	17489
mean	8.04E+04	0.005353	0.029109	0.17364	1631.55912
std	1.55E+05	0.040333	0.105522	0.215523	994.451364
min	3.33E+02	0	0	0	0.4
25%	3.00E+04	0	0	0	869.9
50%	5.00E+04	0	0	0.142857	1481.6
75%	9.00E+04	0	0	0.285714	2257.7
max	1.00E+07	1	1	1	5721.1

4.1.1 Ward Level Distribution of Vulnerability Indicators

Figures 5,6,7,8 and 9 display distribution of five vulnerability indicators among wards of Ratnanagar municipality.

4.1.1.1 Per Capita Income

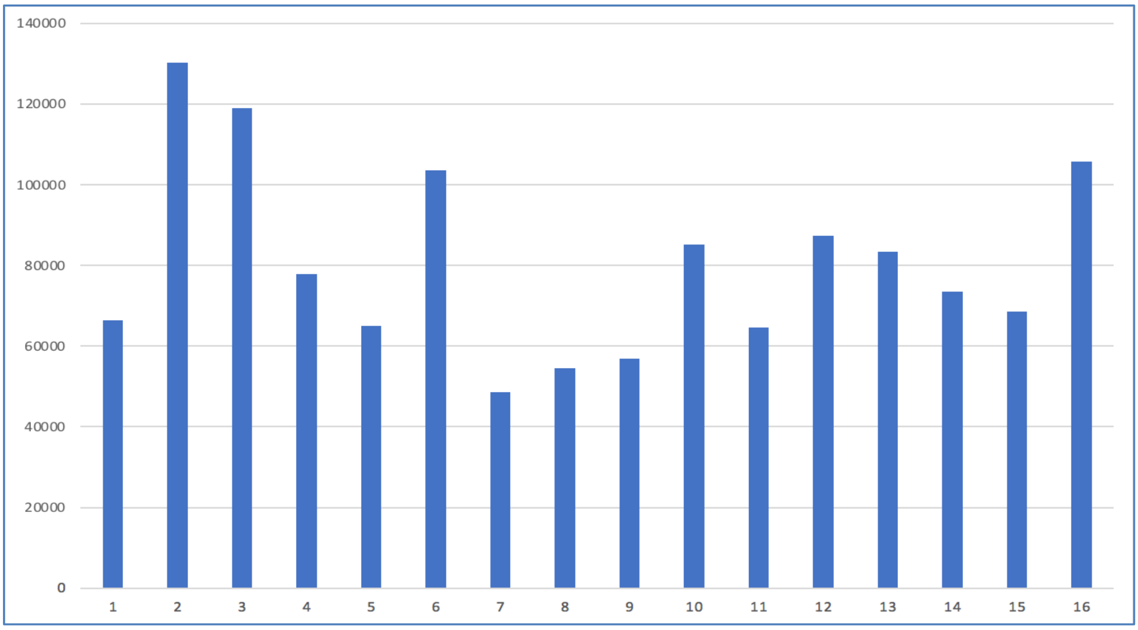


Figure 5. Distribution of average per capita income among wards

Following are the three most and least vulnerable wards in terms of average per capita income:

Table 11. Top 3 most and least vulnerable wards in terms of average per capita income

Most Vulnerable	Average Value (in NPR)	Least Vulnerable	Average Value (in NPR)
7	48606.48622	2	130296.2462
8	54584.33119	3	119085.8051
9	56925.8637	16	105834.308

Wards 7,8 and 9 are considered to be the most remote parts of the municipality. Hence it was no surprise that they were found to be the most vulnerable wards in terms of per capita income.

Wards 2 and 3 lie next to the east west highway. They were found to be the least vulnerable wards in terms of per capita income. Wealthy businessmen are known to live in these wards.

Likewise, ward 16 lies in the north east corner of the municipality. Ex-military people are known to live in this area. It was found to be the third least vulnerable ward in terms of per capita income.

It should be noted that ward 6 comes 4th in the list of least vulnerable wards in terms of per capita income. The famous tourist destination Sauraha lies in this ward. There are a lot of hotels, restaurants, cafes and shops in this ward.

4.1.1.2 Disability Ratio

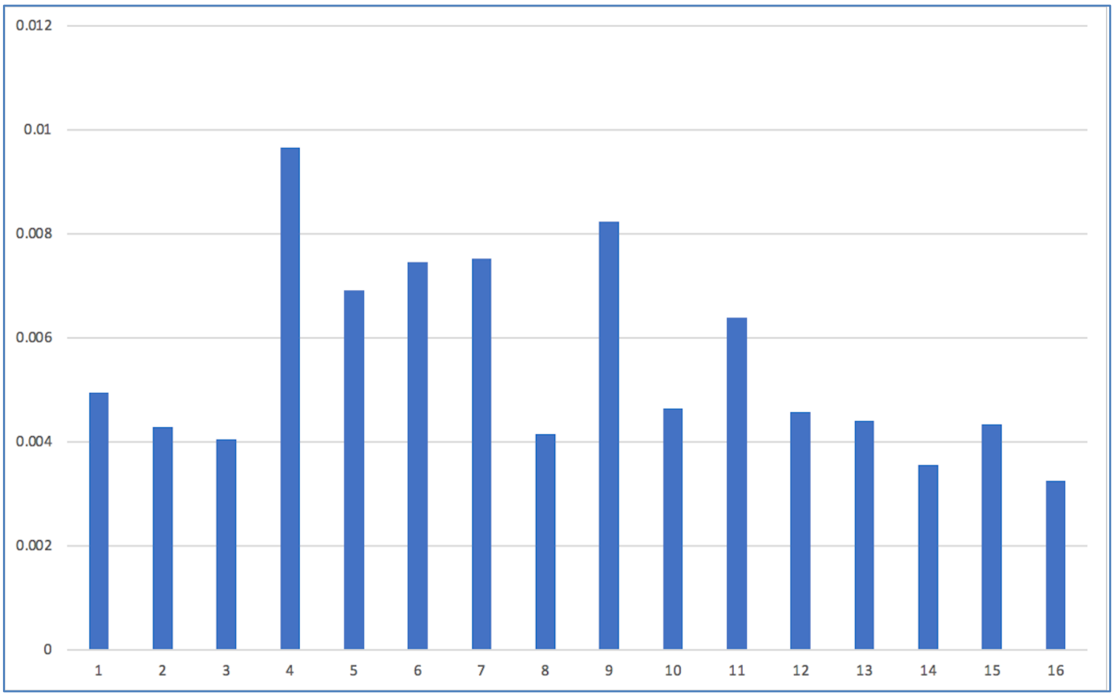


Figure 6. Distribution of average disability ratio among wards

Following are the three most and least vulnerable wards in terms of average disability ratio:

Table 12. Top 3 most and least vulnerable wards in terms of average disability ratio

Most Vulnerable	**Average Value**	**Least Vulnerable**	**Average Value**
4	0.00966659	16	0.003245258
9	0.008233966	14	0.003558353
7	0.007516655	3	0.00404303

Ward 4 was found to be the most vulnerable ward in terms of average disability ratio. Likewise, ward 9 and 7 were found to be the second and third most vulnerable wards in terms of average disability ratio. These wards are also two of the most vulnerable wards in terms of average per capita income.

Ward 16 was found to be the least vulnerable ward in terms of average disability ratio. It is also one of the least vulnerable wards in terms of average per capita income.

4.1.1.3 Ill Ratio

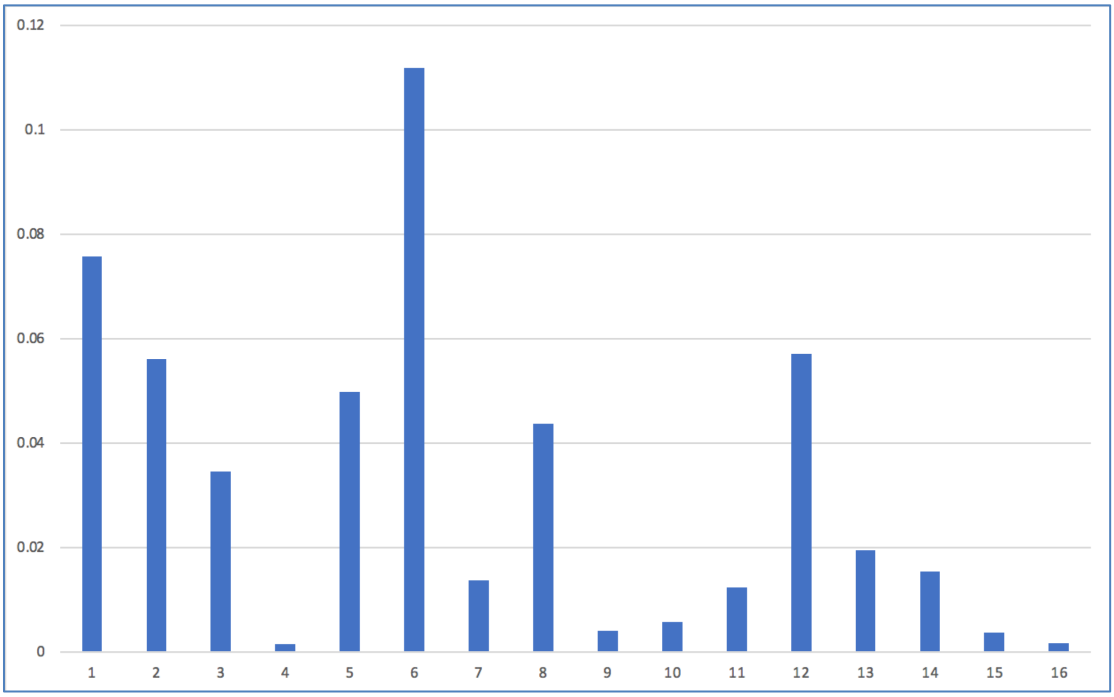


Figure 7. Distribution of average ill ratio among wards

Following wards are the three most and least vulnerable wards in terms of average ill ratio:

Table 13. Top 3 most and least vulnerable wards in terms of average ill ratio

Most Vulnerable	**Average Value**	**Least Vulnerable**	**Average Value**
6	0.111833935	4	0.001441296
1	0.0756634	16	0.001668352
12	0.057166663	15	0.003696602

Ward 6 is the most vulnerable ward of the municipality in terms of average ill ratio. Tharu people form majority of population in this ward. It was observed that ward 6 has exceptionally high vulnerability based on average ill ratio. Ward 1 lies next to the east-west highway and is the second most vulnerable ward in terms of average ill ratio. Likewise, ward 4 is the least vulnerable ward in terms of average ill ratio.

4.1.1.4 Vulnerable Age Ratio

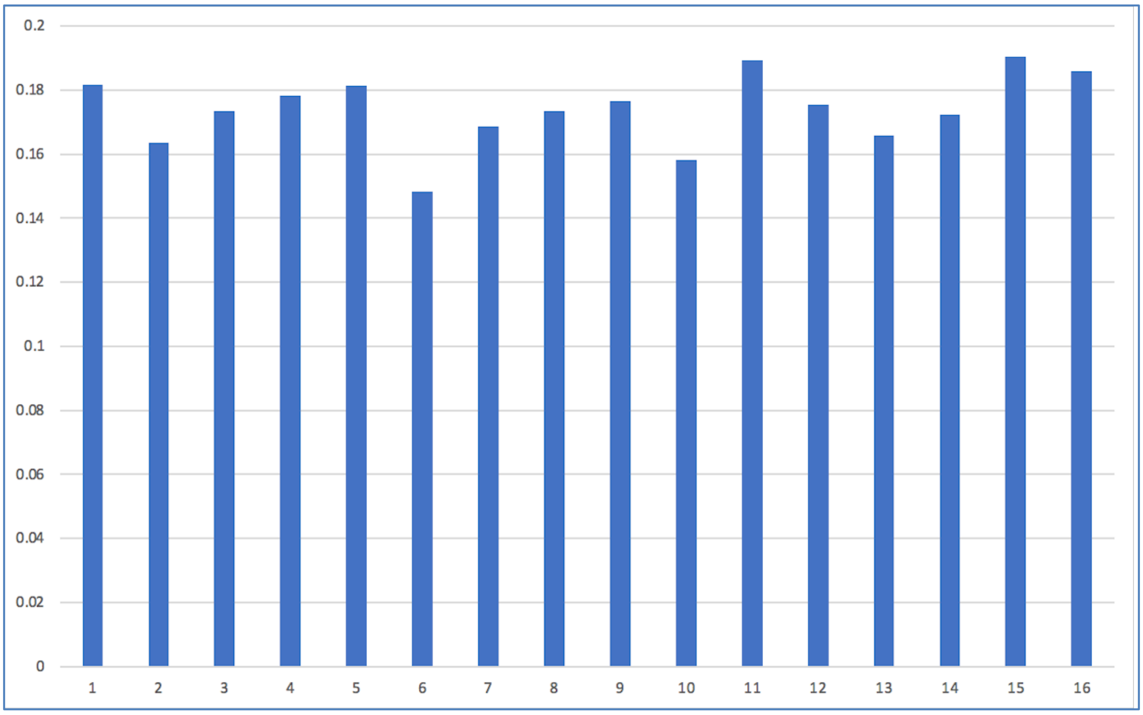


Figure 8. Distribution of average vulnerable age ratio among wards

Following wards are the three most and least vulnerable wards in terms of average vulnerable age ratio:

Table 14. Top 3 most and least vulnerable wards in terms of average vulnerable age ratio

Most Vulnerable	**Average Value**	**Least Vulnerable**	**Average Value**
15	0.190402272	6	0.148238387
11	0.189225869	10	0.158304785
16	0.185975273	2	0.163442207

It was observed that ward 6 is the least vulnerable ward and ward 15 is the most vulnerable ward based on average vulnerable age ratio.

4.1.1.5 Distance to The Nearest Health Service Provider

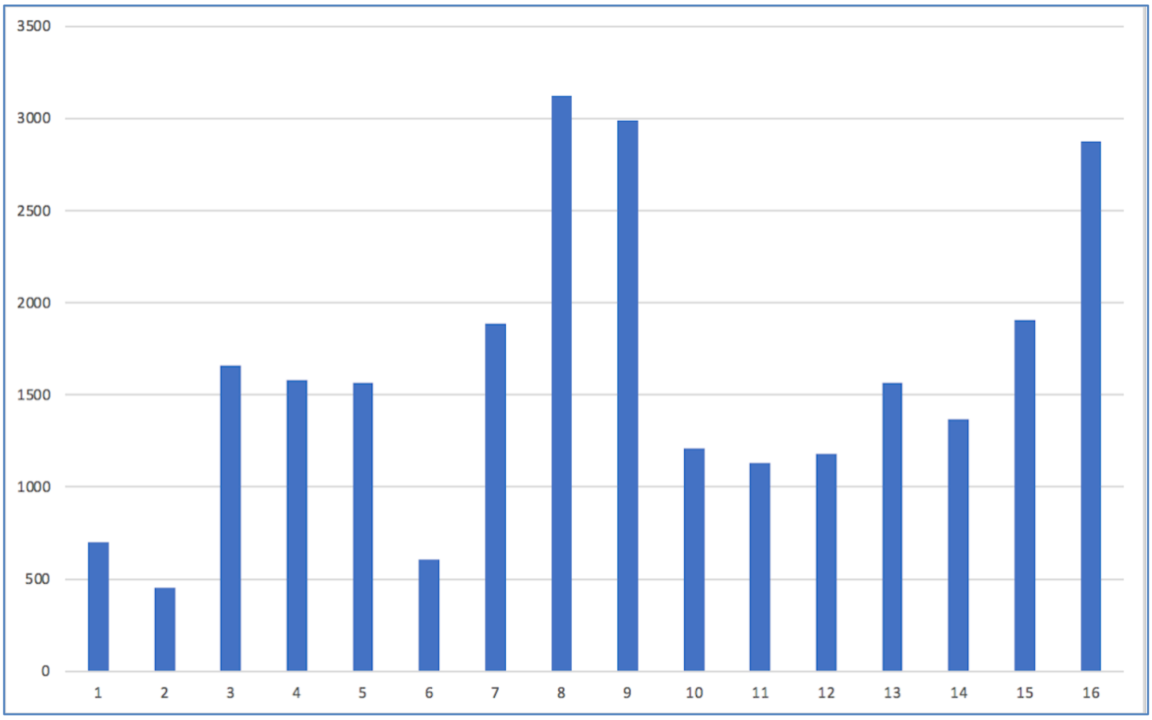


Figure 9. Distribution of average distance to the nearest health service provider among wards

Following wards are the three most and least vulnerable wards in terms of average distance to the nearest health service provider:

Table 15. Top 3 most and least vulnerable wards in terms of average distance to the nearest health service provider

Most Vulnerable	**Average Value (In meters)**	**Least Vulnerable**	**Average Value (In meters)**
8	3123.358184	2	453.2339218
9	2989.829498	6	608.7563975
16	2875.541254	1	699.1201987

Since, wards 8 and 9 belong to the most remote parts of the municipality, they were found to be the most vulnerable wards in terms of average distance to the nearest health service provider. Ward 16 was found to be the third most vulnerable ward in terms of average distance to the nearest health service provider.

Likewise, ward 2 is the least vulnerable ward in terms of distance to the nearest health service provider. Ward 6 is the second least vulnerable ward whereas ward 1 is the third least vulnerable ward in terms of distance to the nearest health service provider.

After performing min-max normalization, the dataset looked as follows:

Table 16. Dataset after normalization

	per_capita_income	**disability_ratio**	**ill_ratio**	**vulnerable_age_ratio**	**nearest_distance**
count	1.75E+04	17489	17489	17489	17489
mean	9.92E-01	0.005353	0.029109	0.17364	0.285133
std	1.55E-02	0.040333	0.105522	0.215523	0.173834
min	-2.22E-16	0	0	0	0
25%	9.91E-01	0	0	0	0.151992
50%	9.95E-01	0	0	0.142857	0.258919

75%	9.97E-01	0	0	0.285714	0.394585
max	1.00E+00	1	1	1	1

Due to min-max normalization, the minimum and maximum values for all the attributes were 0 and 1 respectively.

4.1.2 Vulnerability Mapping

In order to perform vulnerability mapping, K-means++ algorithm was applied. To apply the algorithm on dataset, optimum value of K had to be provided to the algorithm. For that purpose, elbow method was used.

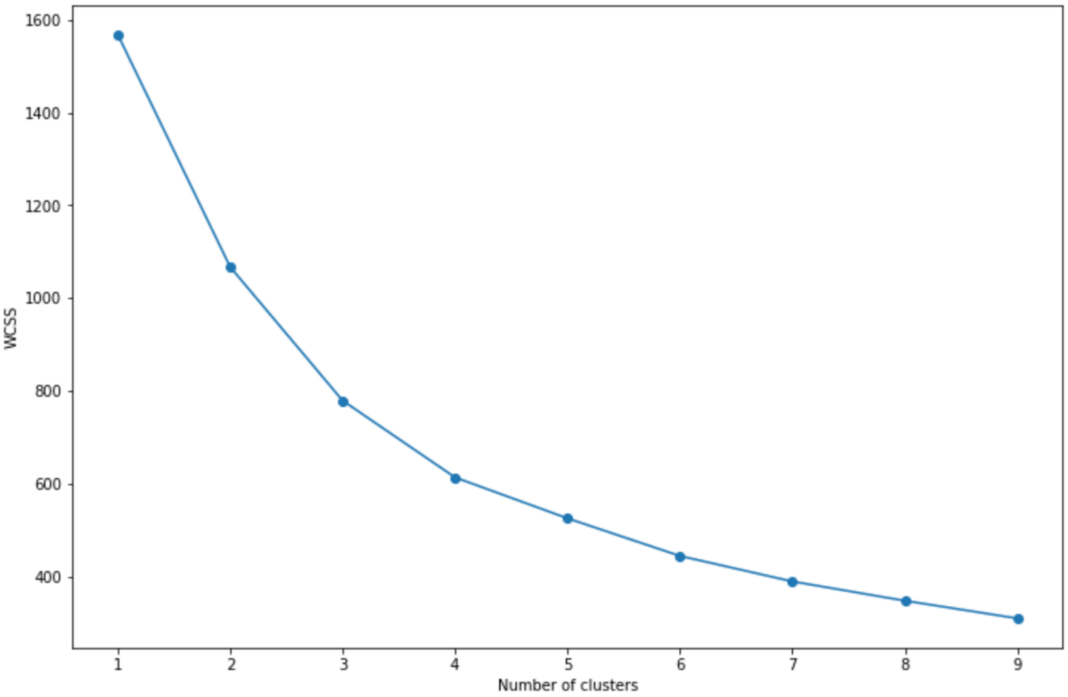


Figure 10. Optimum value of K from elbow method

As seen from the plot, 4 seemed like a good candidate for elbow point of the line. Thus, 4 was chosen as the number of clusters to be formed with K-means++ algorithm.

After performing clustering, the number of households in respective clusters are shown below in figure 11.

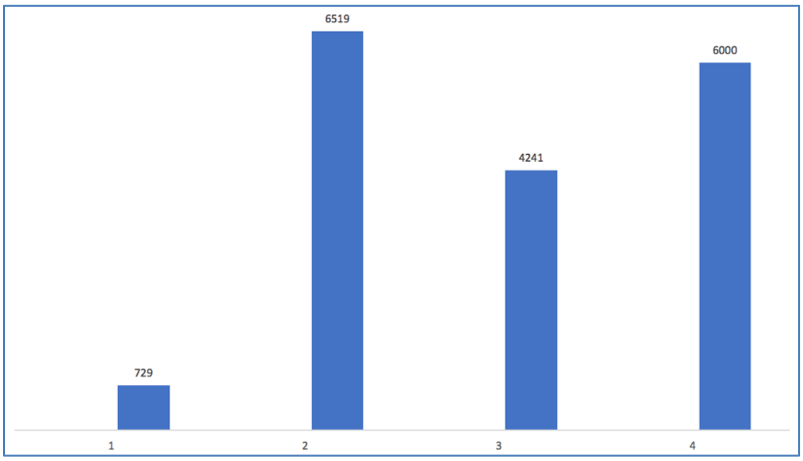


Figure 11. Clusters and respective households' count

Centroid values for the 4 clusters are as follows:

Table 17. Clusters and their centroid values for respective vulnerability indicators

cluster	per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio	nearest_distance
1	0.99216206	0.00902153	0.08902153	0.85854479	0.30874419
2	0.99072689	0.00450545	0.02586313	0.01013786	0.19886775
3	0.99313922	0.00455403	0.01181048	0.12195675	0.52110595
4	0.99255159	0.00639147	0.03757455	0.30450913	0.20918066

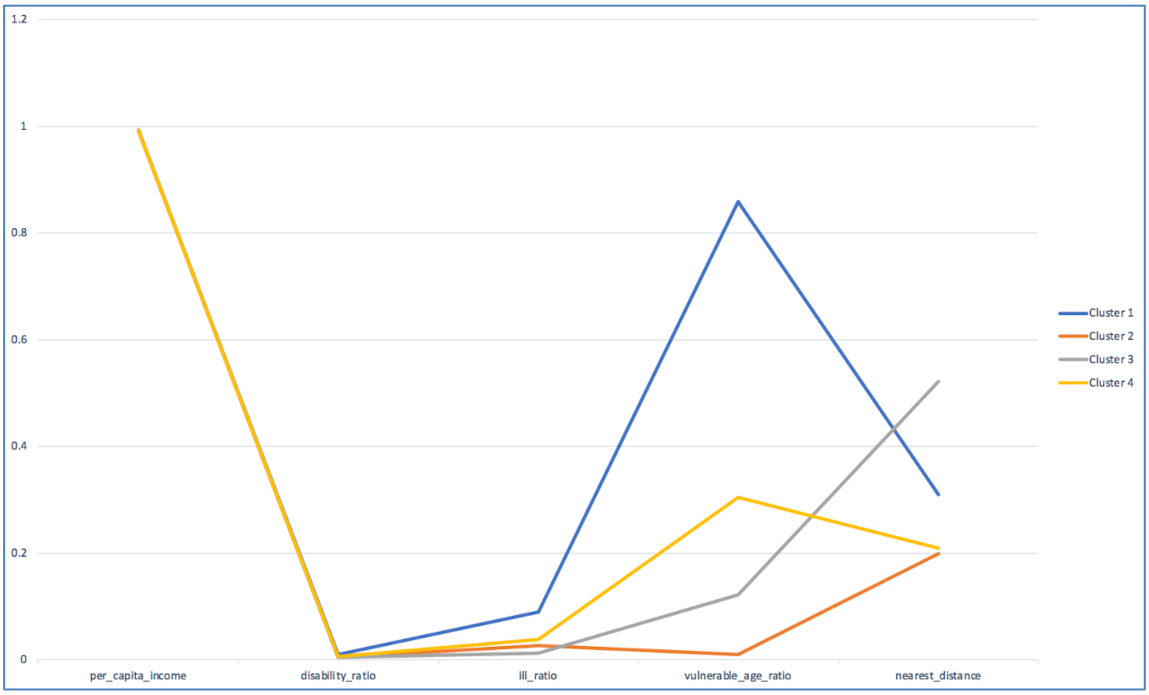


Figure 12. Vulnerability indicators and their centroid values for respective clusters

In the figure above, centroid values of vulnerability indicators for the four clusters have been shown. Each color line represents a particular cluster.

It can be observed that, centroid values of per_capita_income for the four clusters are very close. Likewise, centroid values of vulnerable_age_ratio for the four clusters are very different.

4.1.2.1 Objective weights of vulnerability indicators

The table shown above was used as a decision matrix in order to find the weightage of the individual attributes.

By using entropy method, attribute weights were calculated as follows:

1. **Decision matrix normalization**

In the first step, decision matrix was normalized using Eq. (9).

Table 18. Decision matrix after normalization

per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio	nearest_distance
0.250004314	0.368639732	0.54192304	0.66289292	0.249409927

0.24964268	0.184102846	0.157443114	0.007827565	0.160649475
0.250250538	0.186087797	0.07189688	0.094164298	0.420960144
0.250102469	0.261169625	0.228736966	0.235115217	0.168980453

2. Entropy calculation of vulnerability indicators

In second step, entropy of all the attributes were calculated using Eq. (10).

Table 19. Entropy of vulnerability indicators

per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio	nearest_distance
0.99999971	0.96875861	0.829377689	0.629997746	0.941186752

3. Calculation of degree of diversification of vulnerability indicators

In the third step, degree of diversification of all the attributes were calculated using Eq. (11).

Table 20. Degree of diversification of vulnerability indicators

per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio	nearest_distance
2.89987E-07	0.03124139	0.170622311	0.370002254	0.058813248

4. Calculation of objective weights of attributes

In the fourth and final step of entropy method, objective weights of the attributes were calculated as normalized degree of diversification using Eq. (12).

Table 21. Objective weights of vulnerability indicators

per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio	nearest_distance
4.59801E-07	0.04953608	0.270537274	0.586672404	0.093253782

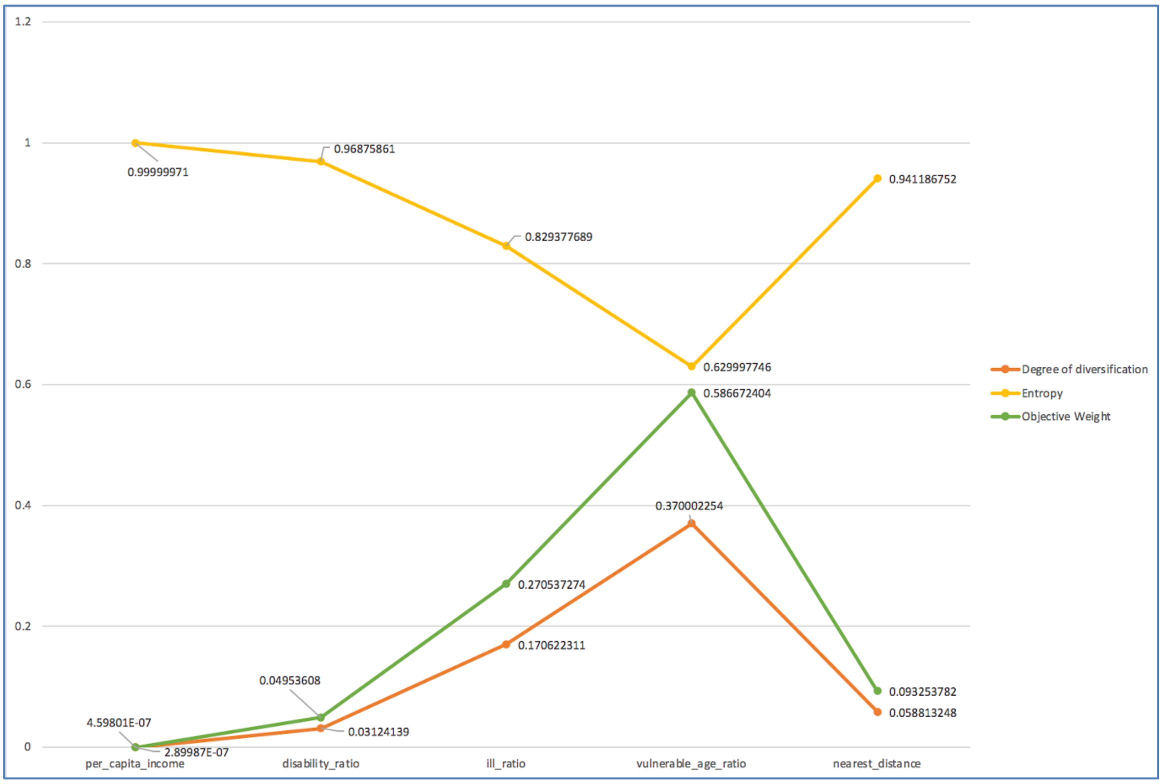


Figure 13. Entropy, degree of diversification and objective weights of vulnerability indicators

From the figure above, it can be observed that per_capita_income has the highest entropy and thus the least degree of diversification. Due to this factor, it has the lowest objective weight. Likewise, vulnerable_age_ratio has the lowest entropy and thus the highest degree of diversification. Due to this, it has the highest objective weight.

4.1.2.2 Cluster vulnerability index

After computing objective weights of all the attributes, Eq. (13) was used to calculate the vulnerability index of all the clusters.

Table 22. Clusters and their respective vulnerability indices

cluster	vulnerability index
1	0.557007089
2	0.031713351

3	0.123564982
4	0.208636371

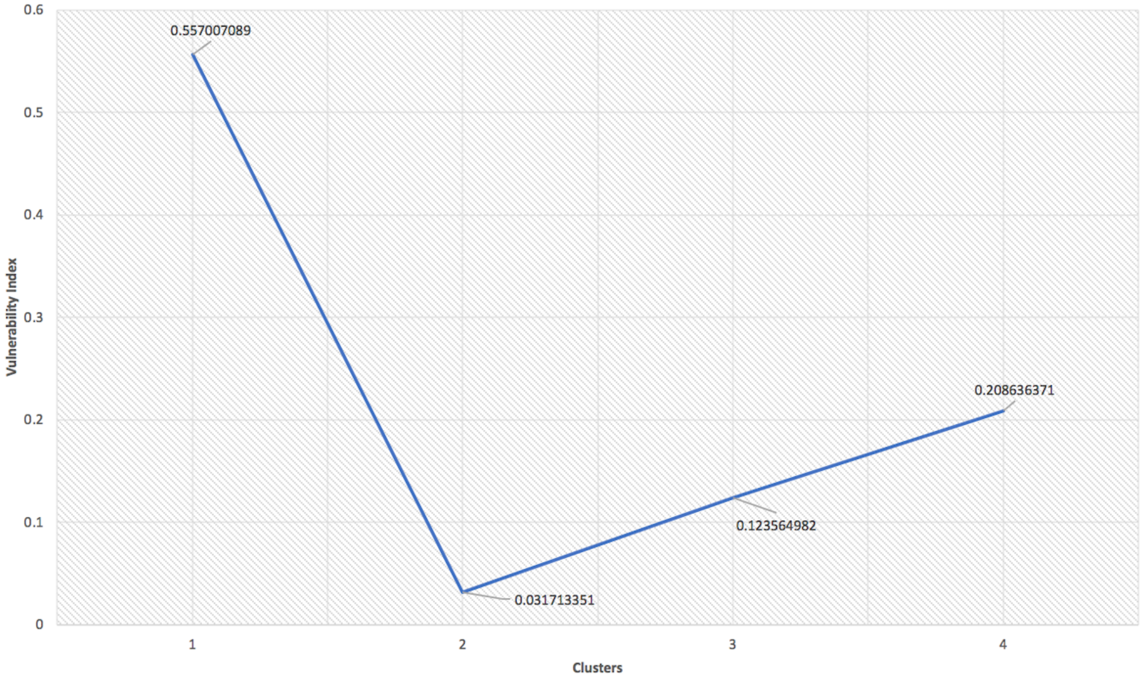


Figure 14. Clusters and their respective vulnerability indices

From the above figure, it can be observed that cluster 1 is the most vulnerable cluster and cluster 2 is the least vulnerable cluster. The reason for it is that, cluster 1 has the highest centroid value for vulnerable_age_ratio attribute. And since this attribute has the highest objective weight, it contributed to the cluster being the most vulnerable cluster.

Likewise, cluster 2 has the lowest value of vulnerable_age_ratio attribute. Due to this, it was the least vulnerable cluster.

4.1.2.3 Ward vulnerability index

Vulnerability indices of the clusters were combined with household counts of individual clusters for a given ward as per Eq. (14) to obtain the vulnerability index of that particular ward as follows:

Table 23. Wards and their respective vulnerability indices

	Number of households belonging to particular cluster				
Ward	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Vulnerability Index**
6	20	452	7	326	0.117211166
10	35	726	45	612	0.123952767
2	23	422	1	347	0.12448245
13	74	868	211	709	0.130365796
14	53	619	86	547	0.133258715
4	50	435	143	345	0.134938338
3	35	343	231	325	0.135678007
1	30	431	0	445	0.136006459
7	25	224	259	233	0.137172163
12	70	699	138	669	0.138190246
11	55	416	67	383	0.143338467
5	47	331	147	313	0.143369557
15	56	363	359	414	0.145503045
9	55	97	930	152	0.14614242
8	47	38	836	81	0.147289732
16	54	55	781	99	0.150638896

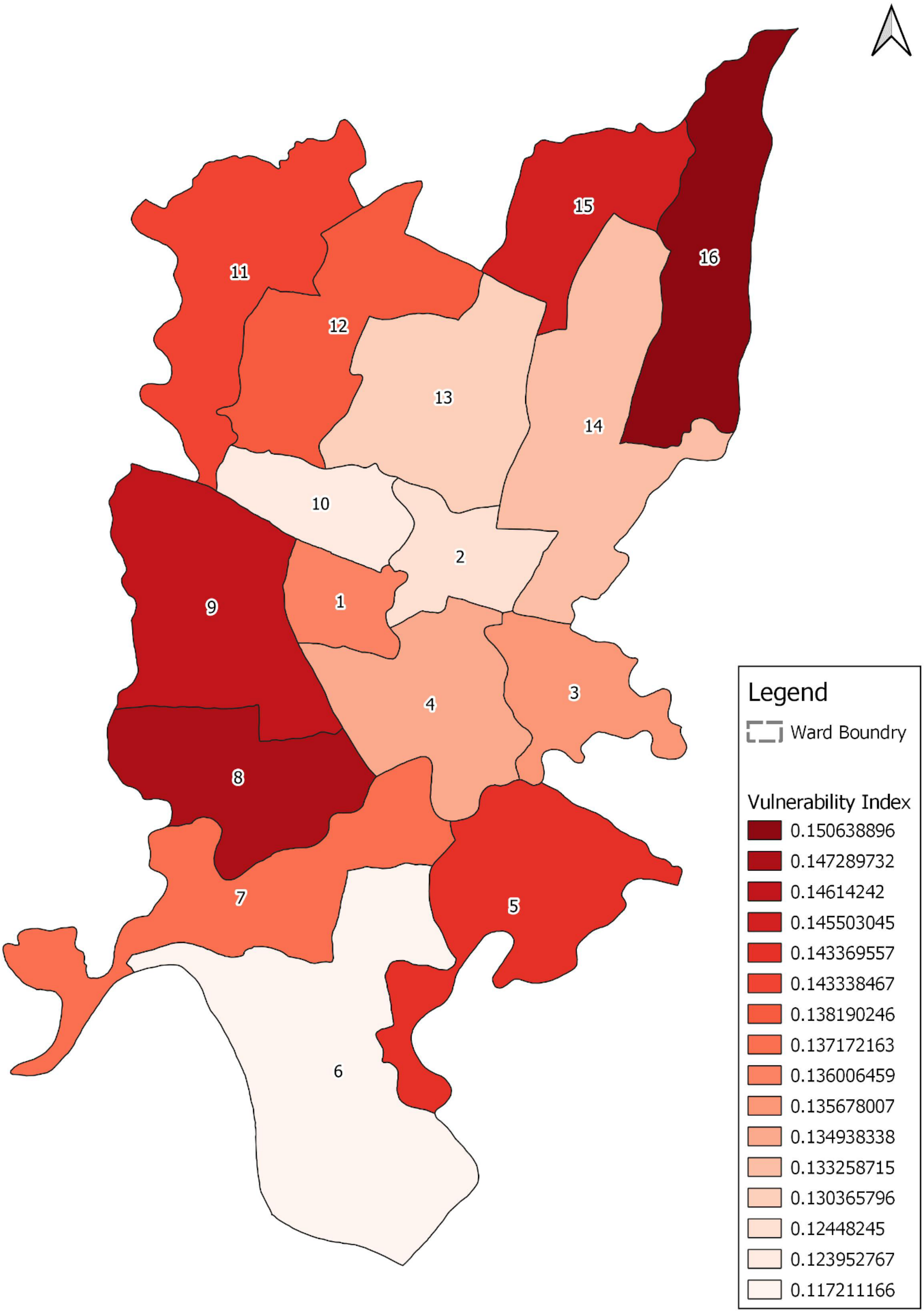


Figure 15. Healthcare vulnerability map of Ratnanagar municipality

4.3 Discussion

As seen from the vulnerability map of the municipality, ward 6 has the lowest vulnerability index whereas ward 16 has the highest vulnerability index. It can be observed that wards which are farther away from the east-west highway have higher vulnerability indices. The exception to this is the ward 6. Except that, the most vulnerable wards are all located near the periphery of the municipality such as wards 16, 8, 9, 15, 5, 11 and 12. Likewise, wards which are close to the east-west highway have lower vulnerability indices such as wards 10, 2 and 4.

Ward 6 was found to be the least vulnerable ward of Ratnanagar municipality. Earlier, ward 6 was found to be the 4th least vulnerable ward in terms of average per capita income. The famous tourist destination Sauraha lies in ward 6. Sauraha lies next to Chitwan National Park, which is visited by a lot of domestic and foreign tourists throughout the year for wildlife exploration. Due to this, there are a lot of hotels, resorts, restaurants, cafes and gift shops in this ward. Higher per capita income of this ward can thus be credited to tourism.

Besides that, it was also found to be the least vulnerable ward in terms of average vulnerable age ratio. In terms of average distance to the nearest health service provider, it was found to be the second least vulnerable ward of the municipality. Due to adequate number of health service providers in the ward, distance to the health service provider was found to be low for this ward.

Despite being at the edge of the municipality, the ward 6 was found least vulnerable, unlike other wards at the periphery, due to sound economic activities, lower average vulnerable age group density and abundance of health service providers.

It should be noted that, vulnerable_age_ratio was the vulnerability indicator with the highest objective weight. Due to the highest degree of diversification of this vulnerability indicator it was assigned the highest objective weight as per entropy method.

Likewise, ward 16 was found to be the most vulnerable ward of the municipality. Pre-clustering analysis had revealed that ward 16 was the third most vulnerable ward in terms of average vulnerable_age_ratio and distance to the nearest health service provider. Thus, this contributed to the ward being the most vulnerable ward in overall. Despite being one of the least vulnerable wards in terms of average per capita income, disability ratio and ill ratio, ward 16 was found to be the most vulnerable ward of the municipality.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

An approach of healthcare vulnerability mapping was presented in this study. Machine learning, geo analytics and entropy methods were combined to evaluate healthcare vulnerability index on ward level. Majority of previous researches showed usage of Principal Component Analysis for dimension reduction in order to calculate vulnerability index. This approach relied on availability of high dimensional data. In case of unavailability of such data, PCA doesn't seem appropriate approach.

This study showed that machine learning algorithm can be effectively used with combination of a weighting method for vulnerability mapping. Application of unsupervised machine learning algorithm made sure that dimensions of vulnerability are visible. Besides showing levels of healthcare vulnerability, the machine learning approach also helped in understanding composition of the vulnerability.

Vulnerability evaluation was done is three stages. In the first stage, households were clustered using k-means++ algorithm. In the second stage, vulnerability indices of individual clusters were evaluated using entropy method. Finally, in the third stage, vulnerability indices of individual wards were computed using vulnerability indices of constituent clusters and their respective part-whole ratios.

In order to apply unsupervised clustering algorithm to the dataset, vulnerability indicators were established and, data preprocessing and feature engineering was done accordingly. Data preprocessing and feature engineering were two of the major tasks performed in this study. As with any machine learning activity, these two constitute the most important steps. The results of machine learning algorithm hugely depend upon the quality of data provided to the algorithm. By applying the steps of relevant attribute selection, data cleaning, normalization and feature engineering, it was made sure that good quality data has been prepared.

Domain knowledge was used to apply proper normalization on different attributes. One of the attributes used was per_capita_income, which had negative correlation with healthcare vulnerability. Thus, negative minmax normalization was applied to this attribute. For rest of the attributes, standard minmax normalization was applied. This made sure that ultimately, all the attributes had positive relationship with the healthcare vulnerability.

Domain knowledge was also used for feature engineering to create an attribute called vulnerable_age_ratio. By considering appropriate age limits for children and elderly and by checking age of individual family members, age wise vulnerable family members were identified.

Finding the distance to the nearest health service provider was the most time-consuming part of the model implementation. As there were 17489 households and 13 health service providers, a total of 2,27,357 distances had to be calculated.

In this study, quality of health service provided was not quantified. Only distance to the health service provider was calculated. Besides that, nature of illness has also not been explored in this research. As part of future work, it is recommended that the quality of health service available be quantified for better analysis. Likewise, nature and severity of illness can also be explored.

Based on contribution of different vulnerability indicators, appropriate interventions can be made for particular wards. The results obtained from this study can be used as baseline data for future researches. As the population of the wards increase, their composition changes and healthcare vulnerability indices change accordingly. The municipality can reevaluate the vulnerability indices and identify the pattern of change in the wards.

REFERENCE

[1] L. Joszt, “5 Vulnerable Populations in Healthcare,” ajmc.com, July 20, 2018. [Online]. Available: https://www.ajmc.com/newsroom/5-vulnerable-populations-in-healthcare. [Accessed Feb. 25, 2020].

[2] D. B. Waisel, “Vulnerable populations in healthcare,” Current Opinion in Anaesthesiology, vol. 26, no. 2, pp. 186-192, April 2013. [Abstract]. Available: https://journals.lww.com/co-anesthesiology/Abstract/2013/04000/Vulnerable_populations_in_healthcare.15.aspx . [Accessed Feb. 25, 2020].

[3] O. A. Abbas, “Comparisons Between Data Clustering Algorithms”, The International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-325, July 2008. Available: https://www.researchgate.net/profile/Osama_Abu_Abbas/publication/220413756_Comparisons_Between_Data_Clustering_Algorithms/links/58d25e57458515b8d28705d5/Comparisons-Between-Data-Clustering-Algorithms.pdf [Accessed Feb. 25, 2020]

[4] S. Cutter, B. Boruff and W. Shirley, "Social Vulnerability to Environmental Hazards", Social Science Quarterly, vol. 84, no. 2, pp. 242-261, 2003. Available: 10.1111/1540-6237.8402002 [Accessed 1 August 2020].

[5] S. Aksha, L. Juran, L. Resler and Y. Zhang, "An Analysis of Social Vulnerability to Natural Hazards in Nepal Using a Modified Social Vulnerability Index", International Journal of Disaster Risk Science, vol. 10, no. 1, pp. 103-116, 2018. Available: 10.1007/s13753-018-0192-7 [Accessed 1 August 2020].

[6] B. de Loyola Hummell, S. Cutter and C. Emrich, "Social Vulnerability to Natural Hazards in Brazil", International Journal of Disaster Risk Science, vol. 7, no. 2, pp. 111-122, 2016. Available: 10.1007/s13753-016-0090-9 [Accessed 1 August 2020].

[7] E. Mavhura, B. Manyena and A. Collins, "An approach for measuring social vulnerability in context: The case of flood hazards in Muzarabani district, Zimbabwe", Geoforum, vol. 86, pp. 103-117, 2017. Available: 10.1016/j.geoforum.2017.09.008 [Accessed 1 August 2020].

[8] O. Žurovec, S. Čadro, and B. Sitaula, “Quantitative Assessment of Vulnerability to Climate Change in Rural Municipalities of Bosnia and Herzegovina,” Sustainability, vol. 9, no. 7, p. 1208, 2017. Available: https://www.mdpi.com/2071-1050/9/7/1208/pdf. [Accessed Jul. 15, 2020].

[9] W. Shi and W. Zeng, "Genetic k-means Clustering Approach for Mapping Human Vulnerability to Chemical Hazards in the Industrialized City: A Case Study of

Shanghai, China", International Journal of Environmental Research and Public Health, vol. 10, no. 6, pp. 2578-2595, 2013. Available: https://www.mdpi.com/1660-4601/10/6/2578/pdf. [Accessed Feb. 25, 2020].

[10] I. Mohamad and D. Usman, "Standardization and Its Effects on k-means Clustering Algorithm", Research Journal of Applied Sciences, Engineering and Technology, vol. 6, no. 17, pp. 3299-3303, 2013. Available: https://pdfs.semanticscholar.org/1d35/2dd5f030589ecfe8910ab1cc0dd320bf600d.pdf. [Accessed Feb. 25, 2020].

[11] C. Yuan and H. Yang, "Research on K-Value Selection Method of k-means Clustering Algorithm", J, vol. 2, no. 2, pp. 226-235, 2019. Available: https://www.mdpi.com/2571-8800/2/2/16/pdf. [Accessed 25 February, 2020].

[12] Central Bureau of Statistics, Government of Nepal, "National Population and Housing Census 2011", 2011. Available: https://unstats.un.org/unsd/demographic-social/census/documents/Nepal/Nepal-Census-2011-Vol1.pdf. [Accessed July 15, 2020].

[13] "Under-five mortality rate (probability of dying by age 5 per 1000 live births)", who.int. [Online]. Available: https://www.who.int/data/gho/indicator-metadata-registry/imr-details/7. [Accessed: 16 July, 2020].

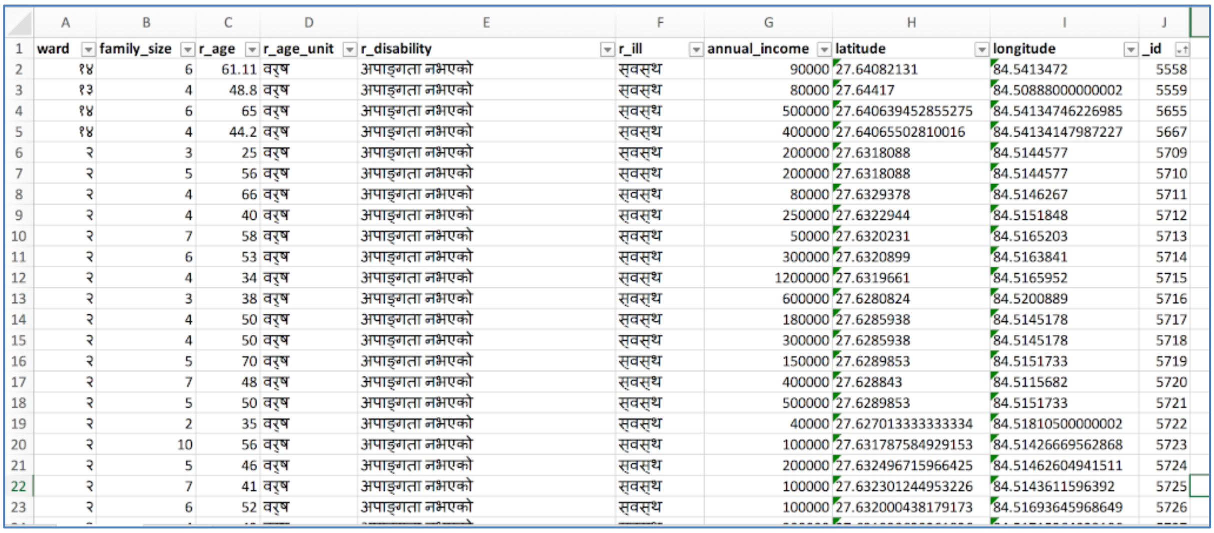
[14] Government of Nepal, ‘Senior Citizens Act’, 2006.

[15] M. Saisana and S. Tarantola, "State-of-the-art report on current methodologies and practices for composite indicator development", 2016. Available: https://www.researchgate.net/profile/Michaela_Saisana/publication/305392511_State-of-the-art_report_on_current_methodologies_and_practices_for_composite_indicator_development/links/578ccb9708ae59aa668146a3/State-of-the-art-report-on-current-methodologies-and-practices-for-composite-indicator-development.pdf. [Accessed: 1 August, 2020]

[16] C.E. Shannon, “A mathematical theory of communication”, Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, 1948. Available: https://pure.mpg.de/rest/items/item_2383162_7/component/file_2456978/content. [Accessed March 5, 2020]

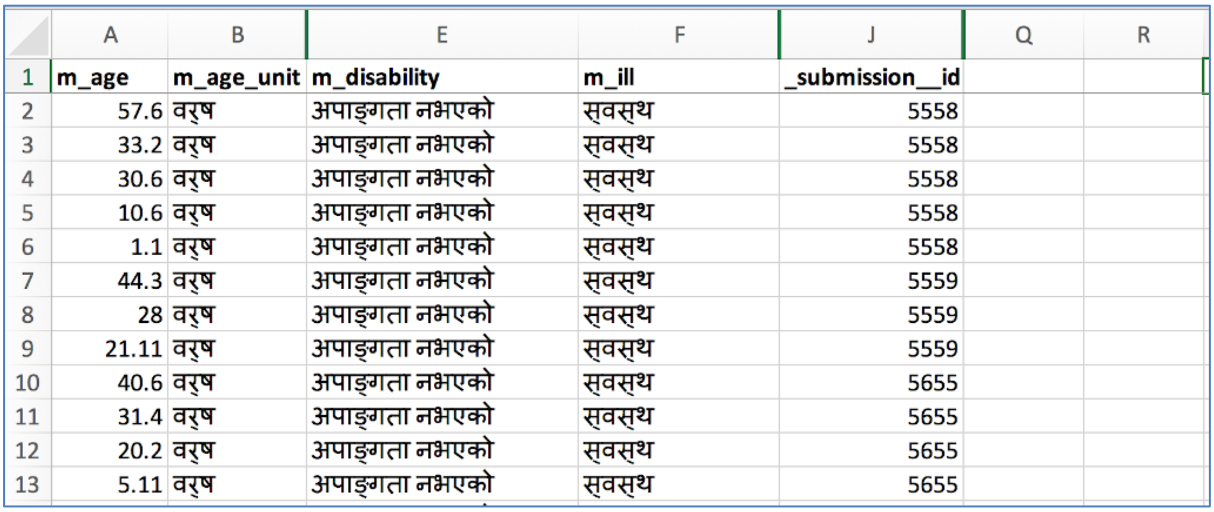
APPENDIX

Datasheets



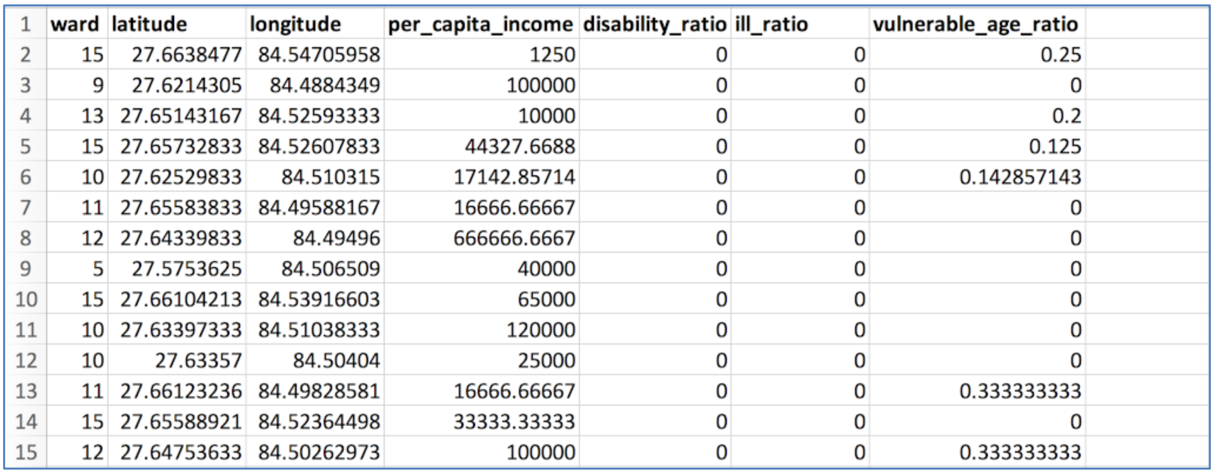
	A	B	C	D	E	F	G	H	I	J
1	ward	family_size	r_age	r_age_unit	r_disability	r_ill	annual_income	latitude	longitude	_id
2	१४	6	61.11	वरुष	अपाङ्गता नभएको	सवस्थ	90000	27.64082131	84.5413472	5558
3	१३	4	48.8	वरुष	अपाङ्गता नभएको	सवस्थ	80000	27.64417	84.50888000000002	5559
4	१४	6	65	वरुष	अपाङ्गता नभएको	सवस्थ	500000	27.640639452855275	84.54134746226985	5655
5	१४	4	44.2	वरुष	अपाङ्गता नभएको	सवस्थ	400000	27.64065502810016	84.54134147987227	5667
6	२	3	25	वरुष	अपाङ्गता नभएको	सवस्थ	200000	27.6318088	84.5144577	5709
7	२	5	56	वरुष	अपाङ्गता नभएको	सवस्थ	200000	27.6318088	84.5144577	5710
8	२	4	66	वरुष	अपाङ्गता नभएको	सवस्थ	80000	27.6329378	84.5146267	5711
9	२	4	40	वरुष	अपाङ्गता नभएको	सवस्थ	250000	27.6322944	84.5151848	5712
10	२	7	58	वरुष	अपाङ्गता नभएको	सवस्थ	50000	27.6320231	84.5165203	5713
11	२	6	53	वरुष	अपाङ्गता नभएको	सवस्थ	300000	27.6320899	84.5163841	5714
12	२	4	34	वरुष	अपाङ्गता नभएको	सवस्थ	1200000	27.6319661	84.5165952	5715
13	२	3	38	वरुष	अपाङ्गता नभएको	सवस्थ	600000	27.6280824	84.5200889	5716
14	२	4	50	वरुष	अपाङ्गता नभएको	सवस्थ	180000	27.6285938	84.5145178	5717
15	२	4	50	वरुष	अपाङ्गता नभएको	सवस्थ	300000	27.6285938	84.5145178	5718
16	२	5	70	वरुष	अपाङ्गता नभएको	सवस्थ	150000	27.6289853	84.5151733	5719
17	२	7	48	वरुष	अपाङ्गता नभएको	सवस्थ	400000	27.628843	84.5115682	5720
18	२	5	50	वरुष	अपाङ्गता नभएको	सवस्थ	500000	27.6289853	84.5151733	5721
19	२	2	35	वरुष	अपाङ्गता नभएको	सवस्थ	40000	27.627013333333334	84.51810500000002	5722
20	२	10	56	वरुष	अपाङ्गता नभएको	सवस्थ	100000	27.631787584929153	84.51426669562868	5723
21	२	5	46	वरुष	अपाङ्गता नभएको	सवस्थ	200000	27.632496715966425	84.51462604941511	5724
22	२	7	41	वरुष	अपाङ्गता नभएको	सवस्थ	100000	27.632301244953226	84.5143611596392	5725
23	२	6	52	वरुष	अपाङ्गता नभएको	सवस्थ	100000	27.632000438179173	84.51693645968649	5726

Figure 16. Respondent worksheet containing respondent and household attributes



	A	B	E	F	J	Q	R
1	m_age	m_age_unit	m_disability	m_ill	_submission__id		
2	57.6	वरुष	अपाङ्गता नभएको	सवस्थ	5558		
3	33.2	वरुष	अपाङ्गता नभएको	सवस्थ	5558		
4	30.6	वरुष	अपाङ्गता नभएको	सवस्थ	5558		
5	10.6	वरुष	अपाङ्गता नभएको	सवस्थ	5558		
6	1.1	वरुष	अपाङ्गता नभएको	सवस्थ	5558		
7	44.3	वरुष	अपाङ्गता नभएको	सवस्थ	5559		
8	28	वरुष	अपाङ्गता नभएको	सवस्थ	5559		
9	21.11	वरुष	अपाङ्गता नभएको	सवस्थ	5559		
10	40.6	वरुष	अपाङ्गता नभएको	सवस्थ	5655		
11	31.4	वरुष	अपाङ्गता नभएको	सवस्थ	5655		
12	20.2	वरुष	अपाङ्गता नभएको	सवस्थ	5655		
13	5.11	वरुष	अपाङ्गता नभएको	सवस्थ	5655		

Figure 17. Member worksheet containing family member attributes



1	ward	latitude	longitude	per_capita_income	disability_ratio	ill_ratio	vulnerable_age_ratio
2	15	27.6638477	84.54705958	1250	0	0	0.25
3	9	27.6214305	84.4884349	100000	0	0	0
4	13	27.65143167	84.52593333	10000	0	0	0.2
5	15	27.65732833	84.52607833	44327.6688	0	0	0.125
6	10	27.62529833	84.510315	17142.85714	0	0	0.142857143
7	11	27.65583833	84.49588167	16666.66667	0	0	0
8	12	27.64339833	84.49496	666666.6667	0	0	0
9	5	27.5753625	84.506509	40000	0	0	0
10	15	27.66104213	84.53916603	65000	0	0	0
11	10	27.63397333	84.51038333	120000	0	0	0
12	10	27.63357	84.50404	25000	0	0	0
13	11	27.66123236	84.49828581	16666.66667	0	0	0.333333333
14	15	27.65588921	84.52364498	33333.33333	0	0	0
15	12	27.64753633	84.50262973	100000	0	0	0.333333333

Figure 18. CSV file containing household attributes

OSRM Server Configuration

1. Download OpenStreetMap data of Nepal from Geofabrik's website at:

   http://download.geofabrik.de/asia/nepal-latest.osm.pbf

2. Navigate to the folder containing the pbf file.

3. Pull the docker image from Docker Hub by issuing following command:

   ```
   docker pull osrm/osrm-backend
   ```

4. Perform preprocessing of OSM data extract by running following command:

   ```
   docker run -t -v "${PWD}:/data" osrm/osrm-backend osrm-extract -p /opt/car.lua
   /data/nepal-latest.osm.pbf
   ```

   This command creates osrm files from the pbf file. Here we used car.lua profile to pre-process the data extract. In our case, it is not important what profile we use since, we are only concerned with distance between two coordinates rather than the time required for travel.

5. Perform partitioning and customization on the newly created osrm file by issuing following commands:

   ```
   docker run -t -v "${PWD}:/data" osrm/osrm-backend osrm-partition /data/nepal-
   latest.osrm
   ```

   ```
   docker run -t -v "${PWD}:/data" osrm/osrm-backend osrm-customize /data/nepal-
   latest.osrm
   ```

6. Run a docker container running the OSRM routing server at port 5000 with following command:

   ```
   docker run -t -i -p 5000:5000 -v "${PWD}:/data" osrm/osrm-backend osrm-routed
   --algorithm mld /data/nepal-latest.osrm
   ```

   Here, argument t assigns pseudo-tty and i runs the container in interactive mode. The algorithm used by the routing server is Multilevel Dijkstra as specified by the argument mld.

   From this point onwards, next time we only have to issue the last command to run the routing server interactively.

Source Code

```
#import all the necessary packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import MinMaxScaler
import requests

#read household and health service provider data
households=pd.read_csv('/Users/apurwa/Downloads/thesis/households-
ratnanagar.csv',header=0)
hsp=pd.read_csv('/Users/apurwa/Downloads/thesis/hospitals-ratnanagar.csv',header=0)

households.describe()

#looping through number of households
x=households[['per_capita_income','latitude','longitude','disability_ratio','ill_ratio','vulnera
ble_age_ratio']]
for ind in x.index:
   #household latitude and longitude
   hh_lat=x['latitude'][ind]
   hh_long=x['longitude'][ind]

   #converting into string format
   hh_lat=str(hh_lat)
   hh_long=str(hh_long)

   #latitude and longitude of first health service provider in csv file
   hsp_lat=hsp['latitude'][0]
   hsp_long=hsp['longitude'][0]

   #converting into string format
   hsp_lat=str(hsp_lat)
   hsp_long=str(hsp_long)

   #constructing URL for HTTP request to OSRM server

url='http://127.0.0.1:5000/route/v1/driving/'+hh_long+','+hh_lat+';'+hsp_long+','+hsp_lat

   #additional parameter
   paramsx={"overview":"false"}

   #making HTTP request against OSRM server
   response=requests.get(url,params=paramsx)

   #response in json
   data=response.json()
```

```
    #storing distance computed in a variable named nearest_distance
    nearest_distance=data['routes'][0]['distance']

    #looping through number of health service providers starting at second instance
    for hind in range(1,len(hsp.index)):

        #health service provider latitude and longitude
        hsp_lat=hsp['latitude'][hind]
        hsp_long=hsp['longitude'][hind]

        #converting into string format
        hsp_lat=str(hsp_lat)
        hsp_long=str(hsp_long)

        #constructing URL for HTTP request to OSRM server

        url='http://127.0.0.1:5000/route/v1/driving/'+hh_long+','+hh_lat+';'+hsp_long+','+hsp
        _lat

        #additional parameter
        paramsx={"overview":"false"}

        #making HTTP request against OSRM server
        response=requests.get(url,params=paramsx)

        #response in json
        data=response.json()

        #storing distance computed in a variable named distance
        distance=data['routes'][0]['distance']

        #checking if newly computed distance is smaller than previously computed distance
        if(distance<nearest_distance):
            nearest_distance=distance

    #distance to the nearest health service provider for a particular household
    x.loc[x.index[ind], 'nearest_distance'] = nearest_distance

x=x[['per_capita_income','disability_ratio','ill_ratio','vulnerable_age_ratio','nearest_distanc
e']]

#Min-max normalization
x_scaled=MinMaxScaler().fit_transform(x)
x_scaled_df=pd.DataFrame(x_scaled)
x_scaled_df.columns=['per_capita_income','disability_ratio','ill_ratio','vulnerable_age_rati
o','nearest_distance']

#since per_capita_income is inversely proportional to vulnerability, we apply negative min
max normalization.
```

```
x_scaled_df.loc[:,'per_capita_income']=1-x_scaled_df.loc[:,'per_capita_income']
x_scaled_df.describe()

#exporting final input to the clustering algorithm as a csv file
x_scaled_df.to_csv("/Users/apurwa/Downloads/thesis/clustering_input.csv",encoding='utf-8')
x_scaled_df.describe()

#elbow method for optimum number of clusters
wcss=[]
for i in range(1,10):
   kmeans=KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init=10,random_state=0)
   kmeans.fit(x_scaled_df)
   wcss.append(kmeans.inertia_)
plt.figure(figsize=(12, 8))
plt.plot(range(1,10),wcss,marker='o')
plt.title('Elbow Method')
plt.xlabel('Number of clusters')
plt.ylabel ('WCSS')
plt.show()

#performing k-means++ clustering algorithm on normalized data
kmeans=KMeans(n_jobs=-1,n_clusters=4,init='k-means++', random_state=0)
y_kmeans=kmeans.fit_predict(x_scaled_df)
x_scaled_df['cluster']=y_kmeans
households['cluster']=y_kmeans
x_scaled_df.describe()

#cluster centroids
kmeans.cluster_centers_

#creating dataframe
clusterframe=pd.DataFrame(data=kmeans.cluster_centers_,columns=["per_capita_income","disability_ratio","ill_ratio","vulnerable_age_ratio","nearest_distance"])

#exporting to csv file
clusterframe.to_csv("/Users/apurwa/Downloads/thesis/centroids_clusters_minmax.csv",encoding='utf-8')

#exporting to csv file
households.to_csv("/Users/apurwa/Downloads/thesis/households_after_clustering_minmax.csv",encoding='utf-8')

#counting number of instances in clusters
x_scaled_df['cluster'].value_counts()
```