BIRZEIT UNIVERSITY

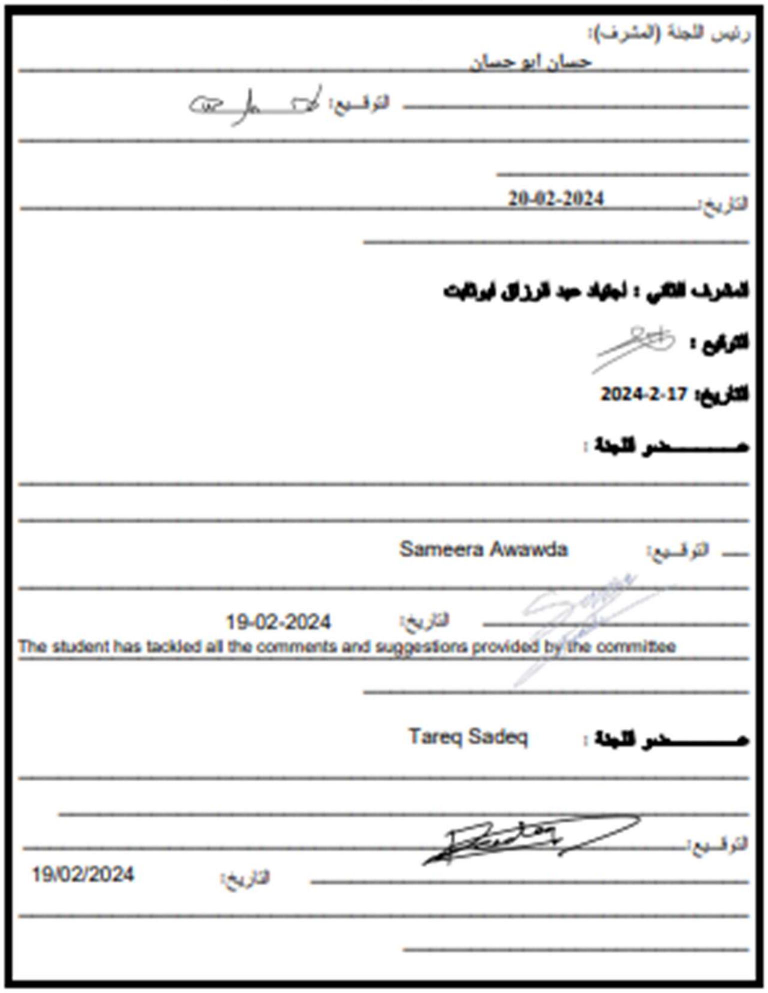
## Faculty of Graduate Studies

Master Program of Applied Statistics and Data Science

# Design a valid and reliable tool to Measure the Vulnerability Index in the West Bank, Palestine

**Prepared by: Nibal Assaly**
**1215152**

**نوقشت هذه الرسالة و اجيزت بتاريخ (19.02.2024) من قبل لجنة المناقشة:**



رئيس اللجنة (المشرف):

حسان ابو حسان

التوقيع:

التاريخ: 20-02-2024

**المشرف الثاني : أ.جهاد عبد الرزاق ابوثابت**

**التوقيع :**

**التاريخ: 2024-2-17**

**عـــــــضو اللجنة :**

التوقيع: Sameera Awawda

التاريخ: 19-02-2024

The student has tackled all the comments and suggestions provided by the committee

**عـــــــضو اللجنة :** Tareq Sadeq

التوقيع:

التاريخ: 19/02/2024

جامعة بيرزيت

BIRZEIT UNIVERSITY

## Faculty of Graduate Studies

**Master Program of Applied Statistics and Data Science**

# Design a valid and reliable tool to Measure the Vulnerability Index in the West Bank, Palestine

**Prepared by: Nibal Assaly**
**1215152**

**First Supervisor: Dr. Hassan Abu Hassan – Applied Statistics and Data Science Program**

**Second Supervisor: Dr. Ijtead Abu Thabet – Applied Statistics and Data Science**

***This Thesis was submitted in partial fulfillment of the requirements for the Master's Degree in Applied Statistics and Data Science from the Faculty of Graduate Studies at Birzeit University, Palestine***

2023-2024

## Acknowledgments

I want to convey my heartfelt appreciation to Dr. Hassan Abu Hassan my first supervisor and Dr. Ijtead Abu Thabet my second supervisor for their consistent support, advice, and direction, which proved invaluable throughout my research and thesis writing. Their extensive knowledge and expertise played a crucial role in the successful completion of this study.

I extend my gratitude to my family and friends who provided unwavering support during my research journey, assisting me in bringing this project to fruition within the stipulated time frame.

# Table of Contents

## Abstract

This research focuses on analyzing the socio-economic vulnerability in the West Bank and mapping the vulnerability patterns. The main objective of this research is to develop a decision-making tool using 2017 socio-economic data collected by the Palestinian Bureau of Statistics (PCBS). In addition, movement restrictions such as the West Bank barrier, checkpoints, and roadblocks, will be part of this analysis, and their impact will be tested using the vulnerability index. The results will be presented spatially on maps in addition to the statistical analyses that will be conducted to show the main reasons for vulnerability.

**Keywords:** Vulnerability, Vulnerability Index, Socio-economic, Indicators, Restrictions of Movement, Multilinear Regression, West Bank, Palestine.

## 1.1 Introduction

In the West Bank, restrictions on movement and economic activity, loss of land and natural resources due to settlements, and violent events have severely affected Palestinian households. The Palestinian household economy is highly sensitive to the conflict and in need of assistance, and unemployment rates (especially among women and youth) are high. Checkpoints and roadblocks, as well as the inability to obtain permits from Israeli authorities, severely limit Palestinians' mobility in the labor market. For farmers and businesses in Area C, which is considered over 60% of the West Bank area, access to markets is an expensive and time-consuming problem that severely limits economic activities and livelihoods. In addition, the land available to Palestinians for farming and livestock is limited due to restrictions on movement and land grabbing by settlers (OCHA, 2013).

The progress of society significantly influences the quality of life, particularly regarding economic growth as a positive factor. Conversely, the present geopolitical landscape and challenges associated with illegal migration greatly diminish the sense of security and consequently impact the overall quality of life negatively (Šoltés, 2016) .

While all Palestinians are vulnerable due to the occupation, some appear to be more vulnerable and systematically more disadvantaged than others. According to the (UN, 2016), Palestinians have been divided into vulnerable groups.

These groups are adolescent girls; Women exposed to gender-based violence; Women-headed households with food insecurity; Children facing barriers to accessing schools; Children in the labor force; Children exposed to violence; Out-of-school children; Adolescents; Elderly; Communities in Area C; Bedouin and pastoralist communities living in Area C; Hebron H2 residents; persons living in the Seam Zones; persons with disabilities; persons in need of urgent medical care; refugees living in abject poverty; refugees in camps; small farmers, non-Bedouin shepherds, fishermen, and the working poor.

UNICEF (2018) mentions that children need humanitarian assistance to receive quality education. The ongoing conflict and violent escalation in the West Bank, including East Jerusalem, as well as restrictions of movement, pose a daily challenge and threat to the realization of children's rights. Violence against children in all its forms is a cause for serious concern, as it affects children's potential for learning and their future. Children are exposed to stress, fear, and intimidation on their way to school in vulnerable areas, where they often have to pass through checkpoints or walk through settlements. Constant exposure to conflict, economic hardship, and increasing poverty contribute to the acceptance of violence as a social norm, which harms children.

The need for humanitarian assistance is growing exponentially, and the environment will become increasingly complex in the coming decade. Climate change, conflict, economic crises, inequality, and pandemics are not new, but these factors affecting the emergency are increasing. They are also interacting and amplifying in unpredictable ways and becoming increasingly irreversible (OCHA, 2022)

For all these reasons, and given the situation on the ground, there was an urgent need to develop a tool to measure the socioeconomic vulnerability of Palestinian communities, based on socioeconomic data from the Palestinian Bureau of Statistics (PCBS).

### 1.2 Problem of the Research

In the West Bank, Palestinians are at risk on several levels. At the individual or household level, vulnerability is determined by internal variables such as social status and income, while at the collective level, vulnerability is determined by external, area- or community-based variables (the natural, economic, sociocultural, and political-institutional environment). People living in politically marginalized and insecure areas are collectively vulnerable because these features of the local environment affect everyone, but the degree to which people are affected depends on their individual or household vulnerability (UN, 2016).

Not all people in an area are equally vulnerable. In the West Bank, it is a series of shocks and processes that affect the lives of Palestinians at different levels and in different ways, changing and shaping the choices that each person in each household must make. In the specific context of the West Bank, vulnerability can be defined as the resilience of an individual or community to withstand and recover from an event, especially from the effects of occupation (UN, 2016).

Therefore, in this study, we consider socioeconomic vulnerability (at the household and locality levels)

### 1.3 Importance of the Research

Socioeconomic vulnerability analysis is crucial in understanding inequalities, identifying populations at risk, guiding policy decisions, evaluating interventions, promoting social justice, and fostering resilience. It serves as a basis for targeted actions and interventions aimed at reducing gaps in societies.

The importance of this research arises from its focus on socioeconomic vulnerability in the West Bank, where it requires a donor response calibrated to this reality. A tool will be created to measure the index of local vulnerability and serve as a decision-making tool. It will help the government and donors assess their priorities according to needs, in addition to conducting data analysis for the long-term consequences of occupation.

### 1.4 Research Objectives

Therefore, the main objectives of the West Bank vulnerability tool are:

- Strengthen the data analysis needed for planning, assistance, and policy in the West Bank, and develop a tool working as a decision-making tool for government or donors needed intervention.
- Reinforce analysis of trends and dynamics in the West Bank, providing solid data to understand the cumulative impacts of the occupation over a longer period.

- Analyze the current obstacles related to occupation, including the West Bank Barrier, closures, and settlements. Demonstrate the connection between these obstacles and the vulnerability index.

### 1.6 Research Questions

Using the results of the vulnerability index and representing it geographically, using the spatial analysis which deals with examinations and interpretations of patterns and relationships. Spatial analysis analyzes the distribution of the vulnerability index for West Bank communities with the occupation obstacles. The same is applied to each indicator used in this study.

1. Representing the unemployment rate geographically with the West Bank Barrier (WBB), Is there a relationship between WBB and the unemployment rate?
2. Is there a relationship between settlements' existence and Vulnerability Index?
3. Is there a relationship between communities in Area C and the Vulnerability Index?

### 1.7 Research Terms and Definitions

- Vulnerability: the fact of being weak and easily hurt physically or emotionally (Oxford Learners Dictionaries, 2023). In this research, it's the diminished capacity of an individual or a community to resist and recover from an event or more specifically from the impact of the occupation. (Laukkonen J, 2009) mentioned that The Vulnerability of individuals and communities to the effects of dangerous events is influenced by more than just the geographical location of their settlements. It is also affected by factors such as the quality of services provided to the settlements, the efficiency and competence of local governments, and the ability of communities to adapt.
- Socio-Economic: based on a social and economic situation (Longman dictionary, 2023). It's the conditions related to social and economic that affect a society (PCBS, 2020).
- Household: all the people living together in a house or flat (Oxford Learners Dictionaries, 2023). According to PCBS, its the number of people who live in a housing unit.
- Indicator: a thing, especially a trend or fact, that indicates the state or level of something. In this research, it's a specific level of information that is related to socio-economic situation (Oxford Languages & Google, 2023). Indicators are statistics used to measure current conditions such as unemployment (PCBS, 2017).
- ArcGIS Pro: a full-featured professional desktop GIS application from Esri. With ArcGIS Pro, you can explore, visualize, and analyze data (ESRI, 2023).

### 1.8 Research Limitations

- Sample Size: Households of Palestinian communities in the West Bank except Jerusalem district.
- Location: West Bank, Palestine
- Time: 2017

- Data Source: Palestinian Bureau of Statistics (PCBS) - Census

### 1.9 Research Challenges and Missing Data

As the data used in this research is mainly from PCBS, and due to occupation, PCBS couldn't have access to Jerusalem (governorate) to do the census (PCBS, 2017). Jerusalem governorate has no data so it will be excluded from the analysis which will affect the overall vulnerability index. Missing data for other districts will be checked and treated by imputation through SPSS if it exists.

# Chapter 2

# Literature Review

## 2.1 Vulnerability Definition

The concept of vulnerability has been a powerful analytical tool for describing states of the ability to harm, powerlessness, and marginality of both physical and social systems, and for guiding normative analysis of actions to enhance well-being through reduction of risk’. Social scientists tend to explain vulnerability as representing the set of socioeconomic factors that determine people’s ability to cope with stress or change (Allen 2003).

According to Laukkonen et al. (2009), the vulnerability of individuals and communities to the consequences of hazardous events is influenced by various factors. These include not only the geographical location of their settlements but also the quality of services provided to those settlements, the effectiveness and competence of local governments, and the adaptive capacity of the communities themselves.

S. Rajesha et al. (2018) say that vulnerability is characterized by the natural inclination of a household to be prone to experiencing harm. It is assessed by combining the effects of natural stressors like severe rainfall, floods, and landslides, with socio-economic factors such as unemployment rates, levels of education, and access to critical services.

According to (UNISDR, 2015), vulnerability is defined as the circumstances shaped by physical, social, economic, and environmental elements or mechanisms, which heighten a community's exposure to the consequences of hazards.

In the West Bank, Palestinians face vulnerability on different levels. On the individual or household level, vulnerability is determined by internal variables such as social status and income, while on the collective level, vulnerability is determined by external, area-related, or community-related variables such as economy, and political situation (UN, 2016).

People living in politically marginalized and insecure areas are collectively vulnerable because these characteristics of the environment affect everybody, but it depends on their individual or household vulnerability to which extent people are affected. Not all people in the area will face equal vulnerability. In the West Bank, it is a series of events that are impacting Palestinian lives at different levels and in different ways. In the specific context of the West Bank, vulnerability can be defined as the resilience capacity of an individual or of a community to resist and recover from an event or more specifically from the impact of the occupation (OCHA, 2022).

## 2.2 Vulnerability in Different Sectors

This section discussed studies examining vulnerability within the health, agricultural, and education sectors, and the subsequent impacts experienced by each sector.

### 2.2.1 Health Sector

According to Raju (2016), an increase in population density leads to a higher dependency on finite resources. Additionally, a higher population density can also potentially result in environmental and health issues.

In a study conducted by Wahyuni (2022) that examined socio-economic indicators about the health sector, particularly focusing on the COVID-19 pandemic, it was found that there is a correlation between higher confirmed COVID-19 cases and a high socio-economic vulnerability index. The study suggests that a high vulnerability index leads to a lower adaptive capacity in addressing health issues, including the challenges posed by the COVID-19 pandemic.

### 2.2.2 Agricultural Sector

In Australia, a comprehensive study by Smith (2014) focused on the agricultural and food sector, specifically encompassing the production of fruits, vegetables, and ornamental plants. The study identified five crucial factors that play a significant role in shaping the socio-economic vulnerability of the agricultural sector. These factors include the percentage of the labor force employed in agriculture, the level of geographic remoteness, the socio-economic advantage/disadvantage of the region, the degree of economic diversity, and the age demographics within the sector. These factors collectively influence the socio-economic resilience and challenges faced by the agricultural industry in Australia.

In 2009, FAO and WFP conducted a study on food security and household vulnerability in Palestine. This study discussed the main reasons for this situation and identified initial actions to reduce the impact of Palestinian household vulnerability.

### 2.2.3 Educational Sector

Studies and research were conducted focusing on the link between education and vulnerability. As mentioned (Wahyuni, 2022), communities experiencing medium to high poverty rates coupled with low levels of education face heightened social vulnerability. Furthermore, (Raju, 2016) mentioned that higher literacy rates can empower communities to diversify their employment opportunities and income sources, thus positively influencing their quality of life and bolstering their resilience against various vulnerabilities. This is because a higher literacy rate corresponds to increased adaptive capacity and awareness, enabling communities to effectively navigate and respond to external pressures.

On the same topic, studies were conducted by (UNICEF, 2018), (ECHO, 2019), and (Gerra et al., 2020). These studies address the impact of the occupation on education. They support initiatives that improve the delivery and quality of education services and build the capacity of humanitarian actors, including first responders, to support proactive and rapid response mechanisms and solutions to minimize disruptions in education.

## 2.3 Global Overview and Expected Solutions

Globally (OCHA, 2022) has published its four-year strategic plan. It identifies six humanitarian needs that will impact high-risk groups: (1) the climate crisis; (2) slow and uneven economic growth; (3) rising inequality; (4) increasing instability, fragility, and conflict; (5) pandemics and disease outbreaks; and (6) a fragmented, competing geopolitical landscape.

A survey was conducted by the UN in collaboration with PCBS on the socioeconomic situation and food security in the West Bank. It found that the most vulnerable groups in the West Bank are among camp residents. Although camp residents are the most represented among aid recipients, the current level of assistance to this group should be maintained to compensate for their hardship (FAO & UN WFP, 2009). While (ECHO, 2019) stated that vulnerable Palestinian communities are:

- those living in Area C and East Jerusalem and Hebron H2, specifically Communities at risk of forced displacement, including Bedouins, and o Communities in the Jerusalem periphery, Block E1, as well as Residents in and around Hebron
- Vulnerable communities with little or no access to basic services
- Households affected by demolition and confiscation of private property and whose livelihoods are at risk

Households living in disadvantaged socioeconomic situations often face poor housing quality, unsafe neighborhoods, inadequate schools, and more stress in their daily lives than other households, with a range of psychological and developmental consequences that can affect their children's development in many ways (Gerra et al., 2020). Furthermore, the dependency rate increases with a higher percentage of non-workers in a district. A higher dependency rate indicates a greater vulnerability in the district. This is because one of the factors that affects people's standard of living is per capita income. With a higher average per capita income, the levels of economic vulnerability are reduced.

Regarding expected solutions, it was mentioned in (ECHO, 2019) that ECHO’s interventions in the West Bank aim to reach the most vulnerable populations with income-generating activities that can be scaled up at the community level. Partners should explore ways to link humanitarian interventions with development interventions to reduce the dependence of target communities on humanitarian assistance. According to a report by (FAO & UN WFP, 2009), greater collaboration between agencies specializing in different areas of intervention is encouraged to ensure that the needs of different target groups are met through appropriately tailored interventions. Joint programming frameworks provide a good platform for integrating the approaches of different organizations.

According to S. Rajesha et al. (2018), the vulnerability index tool has the potential to provide valuable insights into household-level vulnerability, facilitating a deeper understanding of the disparities in vulnerability among households. Such insights can assist decision-makers in devising strategies to mitigate the potential harm that households may face from disruptions or hazards in the future.

A study was conducted by A. Jaafari et al. (2023), highlighting vulnerability due to natural disasters. The study focused on assessing social resilience to natural hazards, specifically concentrating on landslide vulnerability. This was achieved through the modeling and mapping of spatially explicit landslide vulnerability at the county level.
The study mentioned that solutions can be applied by taking actions towards upgrading infrastructure, improving governance, boosting awareness and education, broadening livelihood options, reinforcing social connections, and fostering innovation and knowledge acquisition. The aim is to pinpoint communities and the pertinent elements contributing to decreased resilience against environmental hazards. This enables more precise allocation of

financial resources and implementation of management strategies aimed at mitigating the anticipated adverse socio-economic impacts of natural disasters.

To summarize, the literature review discussed studies that addressed vulnerability in various sectors, whereas this study will encompass most of those sectors such as demographics, employment, access to services, housing, and assets. Additionally, this study focuses on a distinct context, namely occupation, and analyzing the vulnerability index about occupation components such as restriction of movement. S. Rajesha et al. (2018) study overlaps with this study by incorporating a majority of socio-economic indicators, including the unemployment rate, educational attainment, access to water, access to healthcare, illness in households, access to electricity, access to sanitation, and house type owning a vehicle. In addition to that, S. Rajesha et al. (2018) study measures vulnerability index at the level of household, while this study measures it at the level of all community households. Additionally, the study conducted by A. Jaafari et al. (2023) shares similarities with our research in evaluating socioeconomic indicators through expert opinion.

As for the methodology employed, Raju (2016) employed Principal Component Analysis (PCA) to determine the vulnerability index, whereas S. Rajesha et al. (2018) employed Non-Linear Principal Component Analysis (NLPCA) as the indicators used were qualitative and quantitative.
Wahyuni (2022) utilized Spearman rank correlation for their vulnerability index calculations. The methodology adopted in this study consisted of two main steps.
A. Jaafari et al. (2023), the vulnerability of the study area to landslide occurrence was analyzed, quantified, and spatially mapped using the random forest (RF) machine learning technique.
In this study, multilinear regression was used, vulnerability index was found and spatially mapped.

# Chapter 3

# Methodology

### 3.1 Introduction

In this study, statistical and spatial analyses were conducted. Multiple Linear Regression and data mining methods were used to show the added value of this analysis in calculations of the vulnerability index.
Chi-Square test will be conducted to test relationships between occupation features and vulnerability index. In addition, spatial analysis shows how geographically this index is distributed and its link with existing restrictions of movement, West Bank barrier, and settlements.

### 3.2 Data Description

The dataset used for this study consists of 18 socio-economic indicators. The indicators are numerical independent variables which are listed in table 3.1. The data source is the PCBS 2017 census for the West Bank except Jerusalem governorate.

### 3.2.1 Population and Sample

Population: The target population of this study consisted of all communities in the West Bank except all communities in the Jerusalem governorate.
Sample: The sample consisted of all the target population communities.
Analysis Unit: it is the Palestinian community but calculations are based on households within the community.
In addition to that, another geographical data source was the United Nations Office for the Coordination of Humanitarian Affairs organization (OCHA).

### 3.2.2 Population Data

Data is obtained from the PCBS 2017 census (already existing data) which included all Palestinian households in the West Bank (all the Palestinian community). The data is related to 18 socio-economic indicators that are mentioned and explained in section 3.3.

### 3.3 Research Variables

Vulnerability Index: Dependent
Socioeconomic indicators: 18 independent variables.
The selection of socio-economic indicators was based on publications from PCBS regarding socio-economic factors. The first publication emphasized the influence of occupation on the welfare of Palestinian households. This survey was conducted to provide the government and the international community with readily accessible socio-economic indicators for emergency response, aimed at aiding the Palestinian community with their essential needs (PCBS, 2006). The second publication (PCBS, 2016), aimed to establish a vital database concerning the socio-economic factors impacting Palestinian households during the 2014 conflict in the Gaza Strip

Here are the 18 selected indicators:

1. Refugee Status: is a percentage of the population that are registered and non-registered refugees. This calculation is derived from a categorical variable that measures the population with three categories: registered refugee, unregistered refugee, and non-refugee.
2. Educational Attainment: a percentage of the population with low or no education. This calculation is derived from a categorical variable with categories: illiterate, can read and write, elementary, preparatory, secondary, associate diploma, bachelor and above.
3. Household Type: a percentage of the population with composite and extended households. This calculation is derived from a categorical variable with categories: one person, nuclear, composite, and extended household.
4. Household Size: a percentage of the population with a household size above 6. This calculation is derived from a categorical variable with categories: 1-3 persons, 4-5 persons, 6-7 persons, 8-9 persons, 10 and above persons.
5. Type of Housing: is a percentage of housing units that are tents, caravans, barracks, or independent rooms. This calculation is derived from a categorical variable with categories: villa, apartment, house, tent, caravan, barrack, independent room.
6. Tenure of Housing: a percentage of households not owning their housing unit. This calculation is derived from a categorical variable with categories: owned, rented, and others.
7. Housing Density: a percentage of households with 3 or more people per room. This calculation is derived from a categorical variable with categories: less than one person, 1-2 persons, 2-3 persons, and more than 3 people per room.
8. Dependency Ratio: is a percentage of the population aged 0-14 and 60+. It's calculated from a categorical variable with categories aged 0-15 and +60.
9. Unemployment: is a percentage of of economically active population who are unemployed This calculation is derived from a categorical variable with categories: employed, unemployed (economically active: unemployed population except age range 0-14, +60)
10. Female out of labor force: is a percentage of economically inactive females. This calculation is derived from a categorial variable with categories: economically active female employed, economically active female unemployed (females except age range 0-15, +60).
11. Disability: a percentage of the economically inactive population because of disability/aging / illness. This calculation is derived from the categorial variable with categories: disability, aging, or illness.
12. Durable goods: a percentage of households not owning a selection of durable goods. This calculation is derived from the categorical variable with categories: not own stove, not own refrigerator, not own washing machine.
13. Transportation: is a percentage of households not owning a car. This calculation is derived from the categorical variable with categories: own a car, not owning a car.
14. Health insurance: a percentage of the population without health insurance coverage. This calculation is derived from categorical variable with categories: with health insurance, and without health insurance.
15. Drinking water: a percentage of households not connected to the main water network. It's calculated from categorial variables with categories: public water network, water tank, domestic well, public tab, mineral water, and others.

16. Electricity: is a percentage of housing units not connected to the main power grid. This calculation is derived from categorical variable with categories: connected to the main power grid, not connected, private generator.
17. Toilet facility: a percentage of occupied housing units not connected to the main sewage network. This calculation is derived from the categorial variable with categories: wastewater network, tight cesspit, porous cesspit, and others.
18. Internet capacity: is a percentage of households without internet. This calculation is derived from categorical variable with categories: with internet connection, without internet connection.

### 3.4 Data Processing

First, calculation on the PCBS raw data was conducted to find out the percentage of each socio-economic indicator per community. Vulnerability score is calculated for each indicator then the summation of these scores will result in the vulnerability index. Vulnerability score for each indicator is calculated by multiplying its weight by value for each community.
The data calculated will be used for the machine learning algorithm after the vulnerability scores (indicators) are weighted by their importance. According to (Commission, 2020), “When indicators are aggregated into a composite measure, they can be assigned individual weights. This allows the effect or importance of each indicator to be adjusted according to the concept being measured. Weighting methods can be statistical, based on public/expert opinion, or both”. In this study, expert opinion was used. The expert depended first on Maslow’s Pyramid hierarchy of needs and second on some references and studies related to the Palestine context. The expert aligned the indicators with the correspondence hierarchy of the needs.

1. Physiological needs (Base of the Pyramid): these includes necessities and fundamental human needs such as access to essential services, adequate housing, and employment opportunities. Indicators that were selected under this classification are Refugee Status, Educational Attainment, Unemployment, Female out of Labor Force, Drinking Water, and Electricity. This group was given weight 4.
2. Safety and Security: it includes indicators that give a sense of safety and security. Indicators that were selected under this classification are Health Insurance, Tenure of Housing, Dependency Ratio, Household Size, and Toilet Facility. This group was given the weight 3.
3. Love and Belongings: it includes indicators related to social and community-related. Indicators that were selected under this classification are Durable Goods, Transportation, Type of Housing, and Household Type. This group was given the weight 2.
4. Self-Esteem and Actualization: it includes indicators related to self-esteem, recognition personal growth, and fulfilment. The expert chooses these indicators under this classification: Internet Capacity, Disability/Aging/Illness, and Housing Density. This group was given the weight 1.

#### 3.4.1 The Used Software

- SPSS
- R
- ArcGIS Pro: Geographic Information System software

### 3.5 Socio-Economic Indicators

As mentioned, 18 socio-economic indicators were used in this study. About the calculations of indicators, the following paragraph provides a detailed explanation of the computation for each indicator.

1. Refugee Status: is a measure that denotes the legal recognition of an individual as a refugee. It's calculated by having the percentage of refugees about the population. In this study, the sum of registered and unregistered refugees is divided by population.

2. Educational Attainment: it refers to an individual's highest level of education they have acquired. It is commonly measured in terms of completed educational stages, such as primary, secondary, or no education. It gives an idea about the education profile of a community. In this study, it's calculated by adding people with low or no education divided by the population.

3. Household Type: it refers to the composition and structure of a living arrangement, indicating the relationships among individuals living together in a dwelling such as extended family, nuclear family, composite family, and one person household. In this study, it's calculated by adding composite households to extended households divided by total households' number.

4. Household size: Household size refers to the number of individuals who live together and share common living arrangements. In this study, it's calculated by the percentage of households with more than 6 individuals in a community.

5. Dependency Ratio: it's the ratio of the population who are dependent on the working population. It's a young and elderly population. This indicator is calculated by adding the number of people aged between 0-15 years plus a number of people aged over 60 divided by the total population.

6. Unemployment: refers to a situation in which individuals who are capable of working, are actively seeking employment, and are willing to work are unable to find jobs. In this study, it's the percentage of the economically active population who are unemployed.

7. Female out of Labor Force: refers to the portion of the population that consists of economically inactive women. In this study, it's the percentage of women who are e economically inactive in a female population.

8. Disability/Aging/Illness: it refers to the number of people in a community who have disabilities, old, or have illness. In this study, it's the percentage of the economically inactive population because of disability/aging / illness.

9. Type of Housing: it refers to the residential structures that individuals or families occupy. In this study, it's the percentage of housing units that are tents, caravans, barracks, or independent rooms in a community.

10. Tenure of Housing refers to the legal or financial arrangement through which individuals or households occupy and possess their homes. It indicates whether people own their homes or rent them. In this study, it's the percentage of households not owning their housing unit in a community.

11. Housing Density: refers to a number of individuals occupying one room. In this study, it's the percentage of households with 3 or more people per room in a community.

12. Durable Goods: Durable goods refer to products that have a long lifespan and are used over an extended period. Durable goods such as refrigerators, oven, and washing machine. In this study, it's the percentage of households not owning a selection of durable goods (mentioned in table 3.1)

13. Transportation: refers to a population who uses transportation. In this study, it's the percentage of households not owning a car in a community.

14. Health Insurance: refers to a type of coverage that pays for medical and surgical expenses for the insured individual. In this study, it's the percentage of the population without health insurance.

15. Drinking Water: it refers to the source of water for households in a community. The sources could be public water networks, cistern, water trucking. In this study, it's the percentage of households that are not connected to the public water network.

16. Electricity: it refers to the source of electricity for households in a community. It could be a main power grid, generators, or solar panels. In this study, it's the percentage of households not connected to the power grid.

17. Toilet Facility: it refers to the household way of sewage connection. It could be a sewage network or cesspit. In this study, it's the percentage of households who are not connected to sewage network.

18. Internet Capacity: it refers to households with stable connections to the internet. In this study, it's the percentage of households without internet connection.

To apply a regression model analysis, a new attribute will be calculated called vulnerability index, which will be used as the dependent variable to predict vulnerability coefficients of the 18 indicators. The vulnerability index is a summation of vulnerability indices for all indicators.

The calculation of each indicator is explained in table (3.1). The indicator vulnerability index is calculated by multiplying its weight with its ratio.

**Table (3.1) List of indicators and their calculations.**

| Indicator per Community | Weight | Calculations |
|---|---|---|
| Refugee Status | 4 | Number of refugees (registered + unregistered)/community population |

| | | |
|---|---|---|
| Educational Attainment | 4 | Number of people with educational attainment as: Preparatory or elementary or can read & write or illiterate)/community population |
| Household Type | 2 | (composite+extended)/total households in the community. |
| Household Size | 3 | Number of households with size over 6/total households in the community |
| Dependency Ratio | 3 | population aged between 0-15 plus population in the age group 60 or more in the community / total community population |
| Unemployment | 4 | Total unemployed population in the community/number of the economically active population in the community |
| Female Labor Force | 4 | economically inactive females/ total females in the community. |
| Disability/aging/illness | 1 | inactive economic population because of disability, aging, or illness /community population. |
| Type of Housing | 2 | Number of housing units classified as (tents, caravans, barracks, or independent rooms)/total number of occupied housing units in the community |
| Tenure of Housing | 3 | Number of households not owning their housing unit/total number of housing units in the community |
| Housing Density | 1 | number of households with more than 3 people per room/ total housing units in the community |
| Durable Goods | 2 | households not owning stoves, refrigerators, or washing machine/community population. |
| Transportation | 2 | Number of households not owning a car/total housing units in the community |
| Health Insurance | 3 | population without health insurance/ community population. |
| Drinking Water | 4 | Number of households not connected to water network/total community households. |
| Electricity | 4 | Number of households not connected to power grid/total community households |
| Toilet Facility | 3 | Number of occupied housing units not connected to the main sewage network /total occupied units. |
| Internet Capacity | 1 | Number of households without internet connection/ total community households. |

### 3.5 Data Exploration

After the calculations mentioned in the above table (3.1) were done, data was examined for outliers and missing values, then explored by calculating descriptive statistics such as mean, and standard deviation, for all indicators.

# Chapter 4

# Results and Discussion

## 4.1 Introduction

In this chapter, analysis results will be shown and discussed. The main objectives of this study are to predict the vulnerability index using multiple linear regression and to check if this index has an association with occupation features by using statistical tests and geolocation mapping.

To explore the data, descriptive statistics were conducted for the 18 indicators. Table 4.1 below shows these statistics.

**Table (4.1) Descriptive Statistics for the main indicators**

| Indicator | N | Mean | SD | Median | Min | Max | Range | SE |
|---|---|---|---|---|---|---|---|---|
| Dependency Ratio | 521 | 42.88 | 5 | 42.26 | 12.42 | 68.75 | 56.33 | 0.22 |
| Household Size | 521 | 40.3 | 10.06 | 39.32 | 0 | 100 | 100 | 0.44 |
| Educational Attainment | 521 | 50.04 | 9.29 | 48.87 | 13.66 | 100 | 86.34 | 0.41 |
| Refugee Status | 521 | 23.52 | 29.25 | 10.87 | 0 | 100 | 100 | 1.28 |
| Household Type | 521 | 7.91 | 23.92 | 0.25 | 0 | 100 | 100 | 1.05 |
| Unemployment | 521 | 13.62 | 9.43 | 12.16 | 0 | 78.57 | 78.57 | 0.41 |
| Female out of labour Force | 521 | 55.41 | 7.54 | 54.79 | 0 | 100 | 100 | 0.33 |
| Disability | 521 | 4.12 | 2.05 | 4.04 | 0 | 18.18 | 18.18 | 0.09 |
| Tenure of Housing | 521 | 7.23 | 9.59 | 4.97 | 0 | 100 | 100 | 0.42 |
| Housing Density | 521 | 10.1 | 17.29 | 3.95 | 0 | 100 | 100 | 0.76 |
| Type of Housing | 521 | 7.91 | 23.92 | 0.25 | 0 | 100 | 100 | 1.05 |
| Health Insurance | 521 | 33.72 | 18.49 | 33.3 | 0 | 100 | 100 | 0.81 |
| Transportation | 521 | 65.5 | 15.44 | 65.43 | 0 | 100 | 100 | 0.68 |
| Durable Goods | 521 | 4.26 | 9.58 | 1.52 | 0 | 80.48 | 80.48 | 0.42 |
| Drinking Water | 521 | 20.27 | 33.5 | 3.17 | 0 | 100 | 100 | 1.47 |
| Electricity | 521 | 7.2 | 23.99 | 0 | 0 | 100 | 100 | 1.05 |
| Toilet Facility | 521 | 90.27 | 27.13 | 100 | 0 | 100 | 100 | 1.19 |
| Internet Capacity | 521 | 63.47 | 23.35 | 60 | 5.38 | 100 | 94.62 | 1.02 |

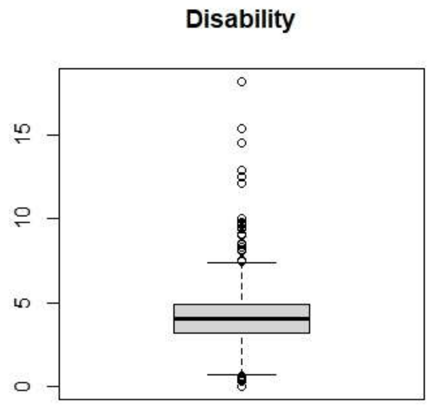
## 4.2 Check for Missing data

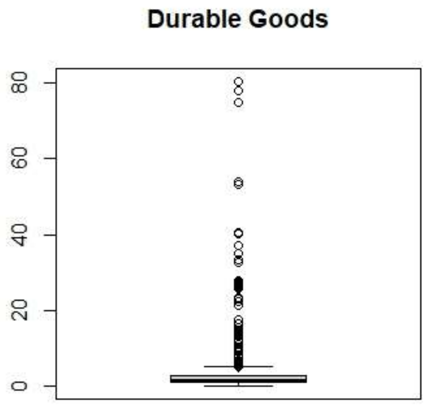
Missing data examination can be applied in SPSS. It can detect missing patterns if existing. If there is missing, the pattern should be examined if it’s completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Understanding the pattern helps in

choosing an imputation technique. Applying the missing data check, there was no missing data in this study.

### 4.3 Check for Outliers

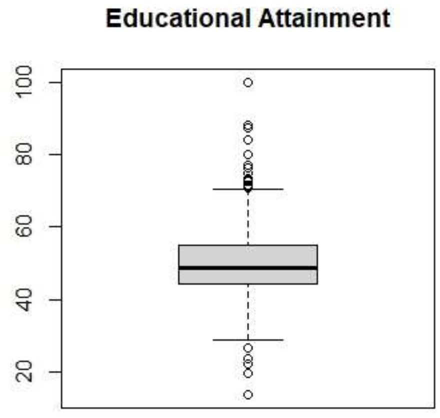
To check for outliers, the Box plots technique will be used. Box plots provide a visual representation of the distribution of data, including potential outliers. Data points outside the box are identified as outliers. Depending on the results, outliers will be examined and treated.

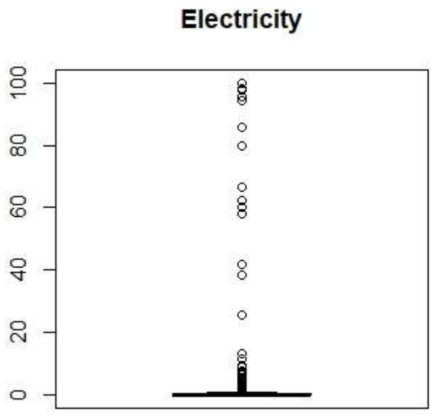




Disability Indicator: outliers in the the end of the data distribution the higher end is more than the lower end of data distribution.

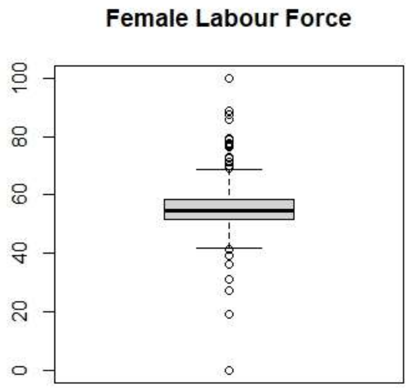
Durable Goods Indicator: outliers lie at the higher end of data distribution.



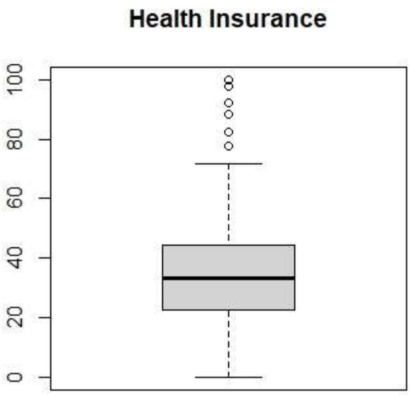


Educational Attainment Indicator: outliers lie between the higher and lower end of the data distribution.

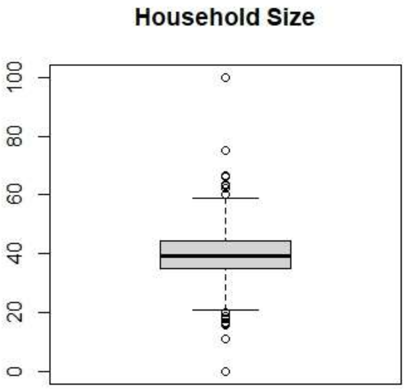
Electricity Indicator: outliers lie in the higher end of data distribution.



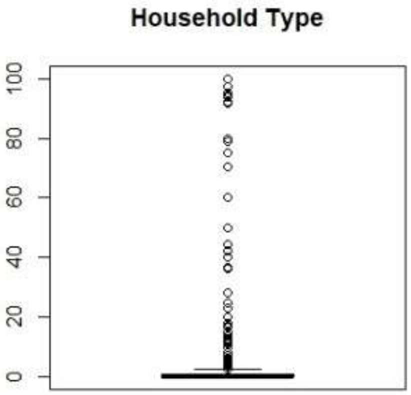
Female Labour Force Indicator: outliers lie between higher and lower end of the data distribution



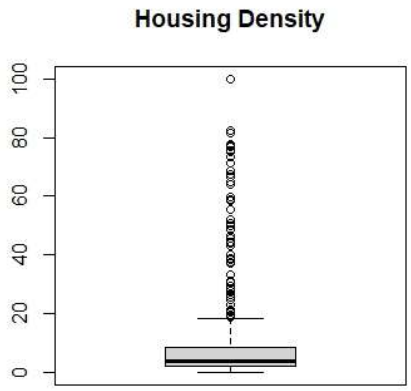
Health Insurance Indicator: outliers lie in the higher end of data distribution



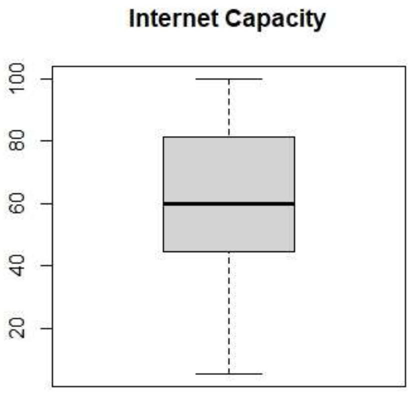
Household size Indicator: outliers lie in between between higher and lower end of the data distribution



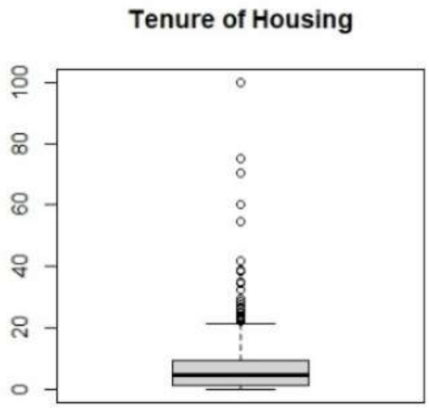
Household Type Indicator: outliers lie in the higher end of data distribution.



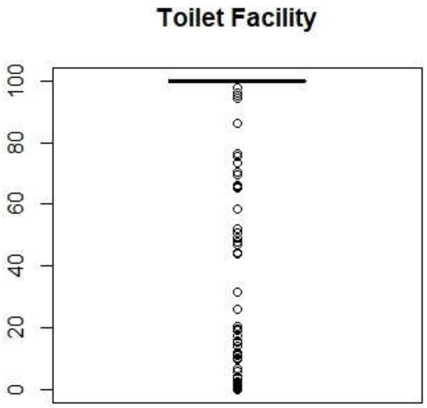
Housing Density Indicator: outliers lie in the higher end of data distribution.



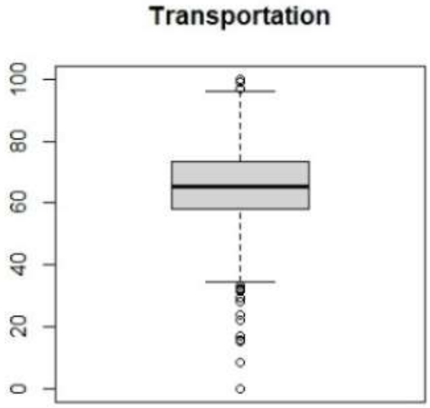
Internet Capacity Indicator: no outliers



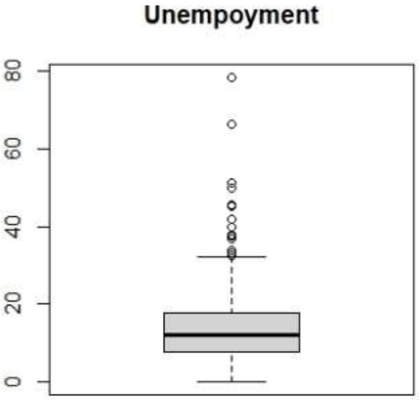
Tenure of Housing Indicator: outliers lie lie in the higher end of data distribution.



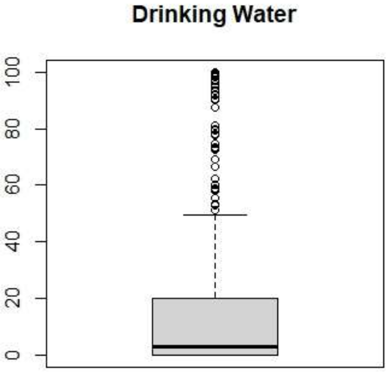
Toilet Facility Indicator: outliers Lie in the lower end of data distribution



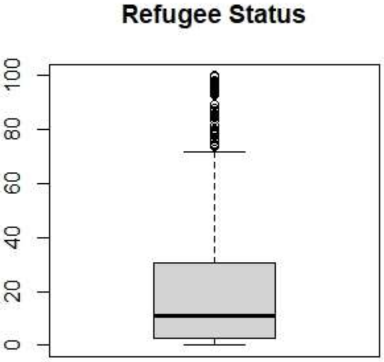
Transportaion Indicator: outliers lie more in the lower end of data distribution.



Unemployment Indicator: outliers lie in the higher end of data distribution



Drinking Water Indicator: outliers lie in the higher end of data distribution.



Refugee Status Indicator: outliers lie in the higher end of data distribution

As detected in the above figures, almost all indicators have outliers. Mahalanobis distance is conducted. It's commonly used in detecting outliers in multivariate data. The rationale behind using Mahalanobis distance for outlier detection is that it accounts for correlations between variables, providing a more accurate measure of distance in multivariate space. Using SPSS, Mahalanobis distance was conducted, any P value less than 0.001 is considered as outlier. table 4.2 shows the list of outliers.

**Table (4.2) list of outliers conducting Mahalanobis distance**

| Governorate | Locality | MAH (MD) | PValue_MD |
|---|---|---:|---:|
| Hebron | Iqtet | 165.23933 | 0.00000 |
| Jenin | Tannin | 160.89458 | 0.00000 |
| Nablus | Khirbet Tana | 181.78358 | 0.00000 |
| Qalqiliya | A'rab Al-Khouleh | 160.95668 | 0.00000 |
| Salfit | Qarawat Bani Hassan | 149.35977 | 0.00000 |
| Salfit | Izbat Abu Adam | 238.36034 | 0.00000 |
| Jericho | An Nabi Musa | 115.22995 | 0.00000 |
| Qalqiliya | A’rab ar Ramadin ash Shamali | 114.71946 | 0.00000 |
| Jenin | Telfit | 113.95167 | 0.00000 |
| Hebron | Khirbet al Kharaba | 101.57464 | 0.00000 |
| Hebron | Imneizil | 99.32774 | 0.00000 |
| Hebron | Khirbet Ghuwein al Fauqa | 95.09649 | 0.00000 |
| Hebron | Khirbet Alrthem | 91.28063 | 0.00000 |
| Salfit | Salfit | 84.75528 | 0.00000 |
| Bethlehem | Khallet A’fana | 81.12014 | 0.00000 |
| Jenin | Fahma al Jadida | 79.74498 | 0.00000 |
| Qalqiliya | A’rab Abu Farda | 79.13595 | 0.00000 |
| Hebron | Khirbet Zanuta | 74.61180 | 0.00000 |
| Hebron | Almefqara | 73.42740 | 0.00000 |
| Hebron | Kafr Jul | 65.60645 | 0.00000 |
| Jenin | Firasin | 65.32983 | 0.00000 |
| Hebron | Haribat an Nabi | 65.32299 | 0.00000 |
| Salfit | Biddya | 65.28182 | 0.00000 |
| Salfit | Deir Ballut | 64.89028 | 0.00000 |
| Hebron | Khirbet ar Rahwa | 64.54236 | 0.00000 |
| Hebron | Maghayir al 'Abeed | 62.93752 | 0.00000 |
| Hebron | Al Maq'ora | 60.71224 | 0.00000 |
| Hebron | Khashem al Karem | 60.45702 | 0.00000 |
| Tubas | Khirbet Tell el Himma | 59.88778 | 0.00000 |
| Hebron | Qawawis | 59.53212 | 0.00000 |
| Nablus | Alttawel and Tall al Khashaba | 59.15372 | 0.00000 |
| Hebron | Khirbet Shuweika | 57.31288 | 0.00000 |
| Salfit | Sarta | 57.12302 | 0.00000 |
| Hebron | Wadi al Kilab | 56.58111 | 0.00000 |
| Jenin | Khirbet al Muntar ash Sharqiya | 55.15452 | 0.00001 |
| Hebron | Sadit athaleh | 54.98223 | 0.00001 |
| Hebron | Khashem Adaraj (Al-Hathaleen) | 54.57427 | 0.00001 |
| Salfit | Masha | 53.19820 | 0.00001 |
| Salfit | Kifl Haris | 53.08181 | 0.00001 |
| Nablus | Furush Beit Dajan | 50.23640 | 0.00004 |
| Jenin | Kherbet Al Hamam | 50.05967 | 0.00004 |
| Hebron | Ar Rakeez | 49.72613 | 0.00005 |
| Salfit | Khirbet Qeis | 47.64017 | 0.00010 |

|  |  |  |  |
|---|---|---|---|
| Tubas | Ibziq | 46.72490 | 0.00013 |
| Hebron | Hamrush | 46.01364 | 0.00017 |
| Salfit | Iskaka | 45.84211 | 0.00018 |
| Tubas | Al Malih | 45.75261 | 0.00019 |
| Salfit | Kafr ad Dik | 44.13232 | 0.00033 |
| Jenin | Khirbet 'Abdallah al Yunis | 43.63177 | 0.00039 |
| Hebron | Edqeqa | 42.57331 | 0.00055 |
| Salfit | Qira | 42.49431 | 0.00057 |
| Jenin | Khirbet Suruj | 42.32530 | 0.00060 |
| Hebron | Birin | 42.23854 | 0.00062 |
| Salfit | Deir Istiya | 41.73404 | 0.00073 |
| Salfit | Bruqin | 41.28564 | 0.00085 |

Table 4.2 shows that 56 communities are considered as outliers. Two models were conducted, with and without outliers to check model specifications. It results with that no major change was detected. Deleting 56 communities means that there will be no vulnerability index for these communities which results in missing values.

In this study we are using real data that represents the real situation on the ground and the true diversity of the population. For these reasons, outliers will not be treated.

### 4.3 Multiple Linear Regression (MLR)

Multiple linear regression is a statistical technique used to model the relationship between a dependent variable and two or more independent variables. It extends the concept of simple linear regression, to include multiple predictors. In our study, this model will be applied between the dependent variable (Vulnerability Index) and other 18 independent variables (18 indicators). The aims of using multiple linear regression are the following:

1. To measure model goodness of fit: to assess and evaluate how much this model fits the data, adjusted R-squared will be used for this purpose.
2. Prediction: the model can be used for the prediction of the dependent variable for future datasets that have the independent variables.
3. To get weights (MLR coefficients) of the indicators that identify the significant impact of each of the independent variables on the dependent variable in terms of strength and direction (positive or negative).

To use this technique, multicollinearity will be examined. Multicollinearity refers to a situation in which two or more predictor variables in a regression model are highly correlated with each other. It can cause problems in statistical analysis, leading to unstable parameter estimates and difficulties in interpreting the results (Daoud, 2017). Variance inflation factor (VIF) is used, which measures the correlation and strength of correlation between the predictor variables in a regression model.

The interpretation of the VIF (Variance Inflation Factor) value is as follows:

- A VIF value of 1 suggests that there is no significant correlation between a particular predictor variable and any other predictor variables within the model.

- A VIF value falling between 1 and 5 indicates a moderate correlation between a specific predictor variable and the other predictor variables in the model. However, this level of correlation is usually not significant enough to necessitate further attention.
- A VIF value exceeding 5 signals the presence of a potentially severe correlation between a particular predictor variable and the other predictor variables in the model. In such cases, the coefficient estimates and p-values in the regression output are likely to be unreliable and should be interpreted with caution (Daoud, 2017).

Using R software, data was split into train and test data then multiple linear regression was conducted, and here is the model summary:

**Table (4.3) Regression Coefficients for Model 1**

| Indicator | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 11.84386 | 0.4926 | 24.044 | < 2e-16 |
| Vul Dependency Ratio | -2.54426 | 0.4452 | -5.715 | 2.68E-08 |
| Vul Disability | 4.53879 | 3.2133 | 1.412 | 0.158857 |
| Vul Drinking Water | 0.99846 | 0.06854 | 14.568 | < 2e-16 |
| Vul Durable Goods | 3.99332 | 0.64558 | 6.186 | 2.06E-09 |
| Vul Educational Attainment | -0.21388 | 0.26445 | -0.809 | 0.419288 |
| Vul Electricity | 0.27644 | 0.13525 | 2.044 | 0.041845 |
| Vul Female Labour Force | -0.93323 | 0.24088 | -3.874 | 0.000132 |
| Vul Health Insurance | 0.72179 | 0.19325 | 3.735 | 0.000225 |
| Vul Household Size | -0.67215 | 0.30003 | -2.24 | 0.025817 |
| Vul Household Type | -0.73413 | 0.88372 | -0.831 | 0.406801 |
| Vul Housing Density | 8.43902 | 0.82247 | 10.261 | < 2e-16 |
| Vul Internet Capacity | 2.97788 | 0.38613 | 7.712 | 1.91E-13 |
| Vul Refugee Status | 0.84846 | 0.0909 | 9.334 | < 2e-16 |
| Vul Tenure of Housing | 0.50538 | 0.2954 | 1.711 | 0.088164 |
| Vul Toilet Facility | 0.91557 | 0.09693 | 9.446 | < 2e-16 |
| Vul Transportation | 1.21988 | 0.28006 | 4.356 | 0.0000183 |
| Vul Type of Housing | 1.86468 | 0.86317 | 2.16 | 0.031557 |
| Vul Unemployment | 0.49607 | 0.17566 | 2.824 | 0.005065 |

Multiple R-squared: **0.923**, Adjusted R-squared: **0.9184**

R-squared and Adjusted R-squared for Model 1 are high, meaning that the model fits the data very well.

VIF test was conducted and below are the results:

**Table (4.4) VIF values for Model 1**

| No. | Indicator | VIF |
|---|---|---|
| 1 | Vul Dependency Ratio | 2.55 |
| 2 | Vul Disability | 1.48 |
| 3 | Vul Drinking Water | 1.68 |
| 4 | Vul Durable Goods | 2.72 |
| 5 | Vul Educational Attainment | 4.22 |
| 6 | Vul Electricity | 5.00 |
| 7 | Vul Female olabour Force | 3.91 |
| 8 | Vul Health Insurance | 2.71 |
| 9 | Vul Household Size | 2.56 |
| 10 | Vul Household Type | 31.79 |
| 11 | Vul Housing Density | 4.25 |
| 12 | Vul Internet Capacity | 2.27 |
| 13 | Vul Refugee Status | 2.65 |
| 14 | Vul Tenure of Housing | 1.30 |
| 15 | Vul Toilet Facility | 1.72 |
| 16 | Vul Transportation | 2.68 |
| 17 | Vul Type of Housing | 33.01 |
| 18 | Vul Unemployment | 2.05 |

Table 4.4 shows VIF values. According to the acceptable values, 16 out of 18 indicators have VIF values between 1 and 5 which indicates a moderate correlation. Two indicators show a high VIF value which are Household Type and Type of Housing which means severe correlation between variables.

Another MLR model will be conducted by removing the indicator with the higher VIF (Type of Housing), thenut re-run the model again.

**Table (4.5) Regression Coefficients for Model 2**

| Indicator | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 11.95096 | 0.49312 | 24.235 | < 2e-16 |
| Vul Dependency Ratio | -2.27303 | 0.42977 | -5.289 | 0.0000 |
| Vul Disability | 4.15487 | 3.22819 | 1.287 | 0.1991 |
| Vul Drinking Water | 0.99825 | 0.06896 | 14.475 | < 2e-16 |
| Vul Durable Goods | 3.9044 | 0.64824 | 6.023 | 0.0000 |
| Vul Educational Attainment | -0.31462 | 0.26191 | -1.201 | 0.2306 |
| Vul Electricity | 0.37679 | 0.12781 | 2.948 | 0.0035 |
| Vul Female out labor Force | -0.98584 | 0.24113 | -4.088 | 0.0001 |
| Vul Health Insurance | 0.69484 | 0.19404 | 3.581 | 0.0004 |
| Vul Household Size | -0.67922 | 0.30186 | -2.25 | 0.0252 |
| Vul Household Type | 1.00762 | 0.36403 | 2.768 | 0.0060 |
| Vul Housing Density | 8.46807 | 0.82743 | 10.234 | < 2e-16 |
| Vul Internet Capacity | 2.93292 | 0.38795 | 7.56 | 0.0000 |

| | | | | |
|---|---|---|---|---|
| Vul Refugee Status | 0.8471 | 0.09146 | 9.262 | < 2e-16 |
| Vul Tenure of Housing | 0.4994 | 0.29721 | 1.68 | 0.0940 |
| Vul Toilet Facility | 0.92365 | 0.09746 | 9.478 | < 2e-16 |
| Vul Transportation | 1.1147 | 0.2775 | 4.017 | 0.0001 |
| Vul Unemployment | 0.59313 | 0.17086 | 3.471 | 0.0006 |

Multiple R-squared: **0.9219**, Adjusted R-squared: **0.9174**

R-squared and Adjusted R-squared for Model 2 decreased a little but are still considered high.

VIF values for Model 2 are shown below:

**Table (4.6) VIF values for Model 2**

| No. | Indicator | VIF |
|---|---|---|
| 1 | Vul Dependency Ratio | 2.35 |
| 2 | Vul Disability | 1.48 |
| 3 | Vul Drinking Water | 1.68 |
| 4 | Vul Durable Goods | 2.71 |
| 5 | Vul Educational Attainment | 4.09 |
| 6 | Vul Electricity | 4.41 |
| 7 | Vul Female out labor Force | 3.87 |
| 8 | Vul Health Insurance | 2.69 |
| 9 | Vul Household Size | 2.56 |
| 10 | Vul Household Type | 5.33 |
| 11 | Vul Housing Density | 4.25 |
| 12 | Vul Internet Capacity | 2.26 |
| 13 | Vul Refugee Status | 2.65 |
| 14 | Vul Tenure of Housing | 1.30 |
| 15 | Vul Toilet Facility | 1.72 |
| 16 | Vul Transportation | 2.60 |
| 17 | Vul Unemployment | 1.91 |

Table 4.6 shows the VIF values for Model 2 which are within the acceptable range (1 -5) except for the indicator Household type, which is higher than 5. As previously done, will do a rerun to the model without this indicator

**Table (4.7) Regression Coefficients for Model 3**

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 11.90113 | 0.49829 | 23.884 | < 2e-16 |
| Vul Dependency Ratio | -2.05672 | 0.42732 | -4.813 | 2.4E-06 |
| Vul Disability | 2.75179 | 3.2237 | 0.854 | 0.39401 |
| Vul Drinking Water | 1.02352 | 0.06912 | 14.808 | < 2e-16 |
| Vul Durable Goods | 4.30733 | 0.63873 | 6.744 | 8.1E-11 |

|  |  |  |  |  |
|---|---|---|---|---|
| Vul Educational Attainment | -0.28336 | 0.26459 | -1.071 | 0.28507 |
| Vul Electricity | 0.53015 | 0.11646 | 4.552 | 7.8E-06 |
| Vul Female Labour Force | -0.95939 | 0.24363 | -3.938 | 0.0001 |
| Vul Health Insurance | 0.68211 | 0.19615 | 3.478 | 0.00058 |
| Vul Household Size | -0.95032 | 0.28872 | -3.292 | 0.00112 |
| Vul Housing Density | 9.79302 | 0.68244 | 14.35 | < 2e-16 |
| Vul Internet Capacity | 2.91269 | 0.39221 | 7.426 | 1.2E-12 |
| Vul Refugee Status | 0.84779 | 0.09248 | 9.168 | < 2e-16 |
| Vul Tenure of Housing | 0.48687 | 0.30049 | 1.62 | 0.10624 |
| Vul Toilet Facility | 0.94137 | 0.09833 | 9.574 | < 2e-16 |
| Vul Transportation | 1.08702 | 0.28041 | 3.877 | 0.00013 |
| Vul Unemployment | 0.56618 | 0.17249 | 3.282 | 0.00115 |

Multiple R-squared: **0.9199**, Adjusted R-squared: **0.9155**

R-squared and Adjusted R-squared for Model 3 decreased a little but are still considered high.

VIF values for Model 3 are shown below:

**Table (4.8) VIF values for Model 3**

| No. | Indicator | VIF |
|---|---|---|
| 1 | Vul Dependency Ratio | 2.27 |
| 2 | Vul Disability | 1.44 |
| 3 | Vul Drinking Water | 1.65 |
| 4 | Vul Durable Goods | 2.58 |
| 5 | Vul Educational Attainment | 4.08 |
| 6 | Vul Electricity | 3.58 |
| 7 | Vul Female out labor Force | 3.86 |
| 8 | Vul Health Insurance | 2.69 |
| 9 | Vul Household Size | 2.29 |
| 10 | Vul Housing Density | 2.82 |
| 11 | Vul Internet Capacity | 2.26 |
| 12 | Vul Refugee Status | 2.65 |
| 13 | Vul Tenure of Housing | 1.30 |
| 14 | Vul Toilet Facility | 1.71 |
| 15 | Vul Transportation | 2.60 |
| 16 | Vul Unemployment | 1.91 |

Table 4.8 shows the VIF values for Model 3 which are within the acceptable range (1 -5).

**Homoscedasticity:** to check for Homoscedasticity, figures 4:1 and 4.2 show a scatter plot of residuals vs fitted or predicted values. If the points are equally scattered above and below the horizontal line, it indicates that the assumption of homoscedasticity is met (Gujarati, 2021). It can be said that for our model homoscedasticity is met.

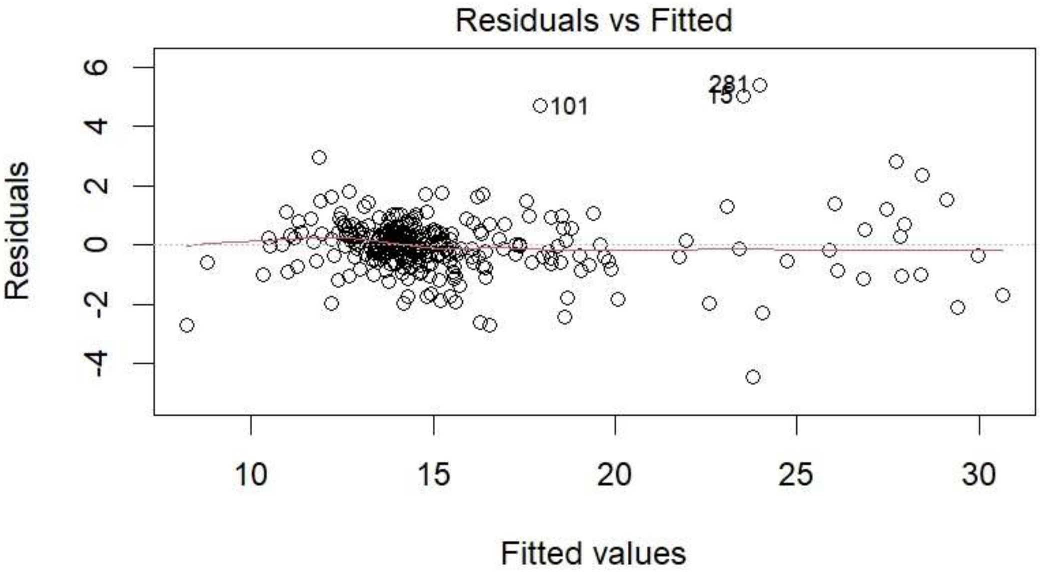


**Figure 4:1 Residuals VS Fitted**

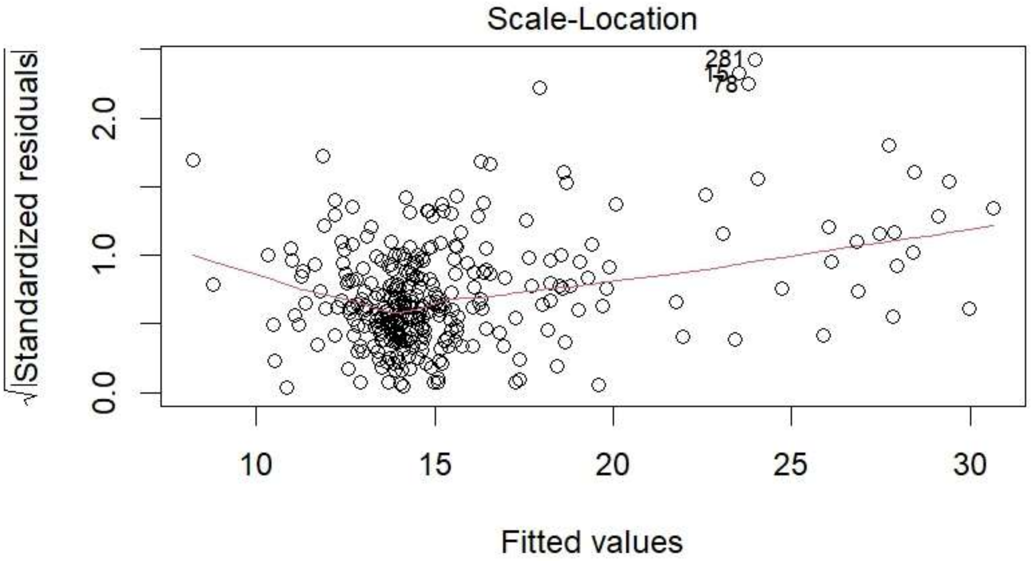


**Figure 4:2 Scale Location**

**Normality:** In a Q-Q residuals plot, if the points fall approximately along a straight diagonal line, it suggests that the residuals closely follow a normal distribution (Gujarati, 2021). In our case, it can be seen that there are deviations from the diagonal line which indicate that the residuals do not follow a normal distribution.

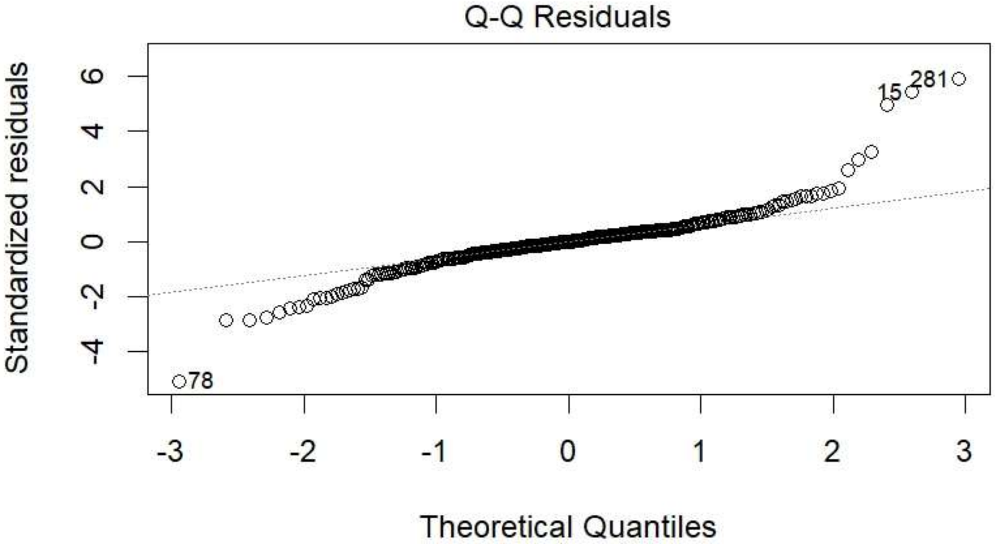


**Figure 4:3 Q-Q Plot for Residuals**

For more investigation on data normality, the Shapiro-Wilk test was conducted for model residuals. Hypotheses for the Shapiro test are:

$H_0$: data is normally distributed.
$H_a$: data is not normally distributed.

data: residuals (Model 3)
W = 0.90677, p-value = 5.149e-13

As the P value is <0.05, we reject $H_0$ which concludes to that the data is not normally distributed.
In this case, we need to apply data transformation. Log transformation was applied to the data.

The same procedure was applied to the log data. Data was split into training and testing sets. MLR was conducted. Applying VIF calculations, the same procedure that was applied to the 3 Models, deleting the high VIF indicator and rerunning the model. It ends with having 15 indicators out of 18.

Table 4.9 shows the final log model coefficients.

**Table (4.9) Regression Coefficients for Log Model**

| Indicator | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---:|---:|---:|---:|
| Intercept | 25.8205 | 0.9455 | 27.31 | < 2e-16 |
| Log Vul Dependency Ratio | -4.853 | 1.6473 | -2.946 | 0.00347 |
| Log Vul Disability | 0.4923 | 0.5156 | 0.955 | 0.34044 |
| Log Vul Drinking Water | 1.6132 | 0.1712 | 9.421 | < 2e-16 |
| Log Vul Durable Goods | 1.8754 | 0.3478 | 5.392 | 0.00000 |
| Log Vul Educational Attainment | -0.9462 | 0.6869 | -1.378 | 0.16938 |

| | | | | |
|---|---|---|---|---|
| Log Vul Electricity | 1.1065 | 0.2375 | 4.658 | 0.00000 |
| Log Vul Health Insurance | -0.8824 | 0.3577 | -2.467 | 0.01420 |
| Log Vul Household Size | -1.8439 | 1.0235 | -1.801 | 0.07264 |
| Log Vul Household Type | 0.495 | 0.2376 | 2.083 | 0.03811 |
| Log Vul Housing Density | 1.7236 | 0.3394 | 5.079 | 0.00000 |
| Log Vul Internet Capacity | 3.6096 | 0.7503 | 4.811 | 0.00000 |
| Log Vul Refugee Status | -0.0185 | 0.2305 | -0.08 | 0.93607 |
| Log Vul Tenure of Housing | -0.1559 | 0.3019 | -0.516 | 0.60601 |
| Log Vul Toilet Facility | 0.7479 | 0.2801 | 2.67 | 0.00801 |
| Log Vul Unemployment | 0.9438 | 0.3552 | 2.657 | 0.00830 |

Multiple R-squared: **0.777**, Adjusted R-squared: **0.7**66

R-squared and Adjusted R-squared for the Log Model decreased from the 90s to almost 80's. Still, the model is a good fit for the data.

Shapiro-Wilk test was conducted for the Log Model,

data: residuals (Log Model)
W = 0.994, p-value = **0.251**

as P value > 0.05, it implies to reject $H_o$, so the data is normally distributed. Plotting the Log Model:

**Homoscedasticity:** Figures 4:4 and 4.5 show scatter plots of residuals vs fitted or predicted values. If the points are equally scattered above and below the horizontal line, it indicates that the assumption of homoscedasticity is met. It can be said that for our Log Model homoscedasticity is met

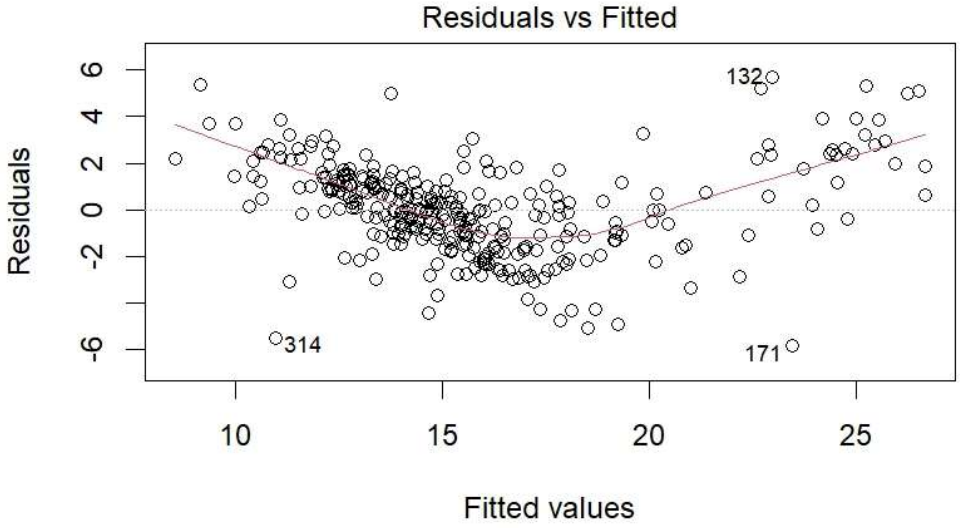


**Figure: 4.4 Residuals VS Fitted**

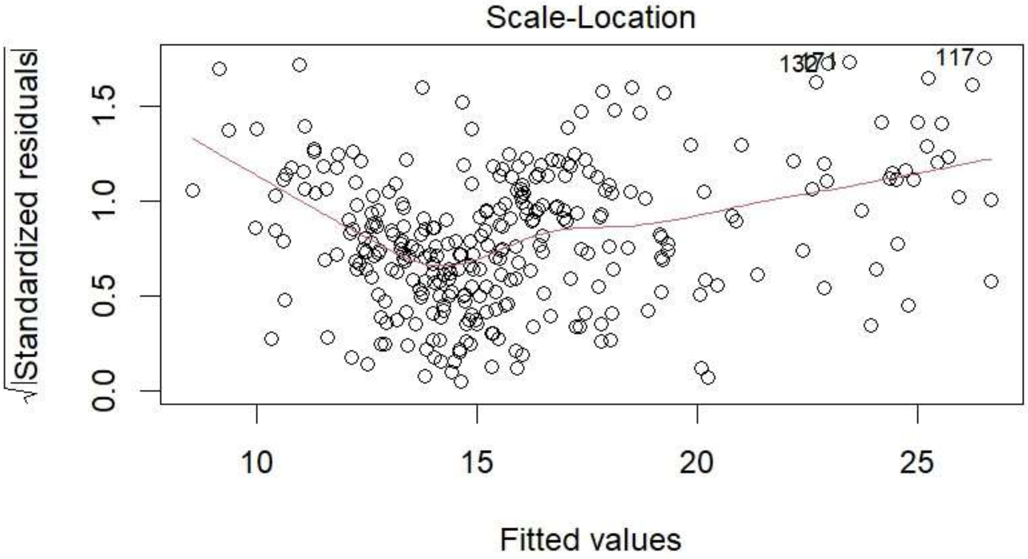


**Figure: 4.5 Scale Location**

**Normality:** figure 4.6 shows the Q-Q residuals plot for the Log Model. It can be seen that points fall approximately along a straight diagonal line, which suggests that the residuals closely follow a normal distribution by conducting the Log transformation.

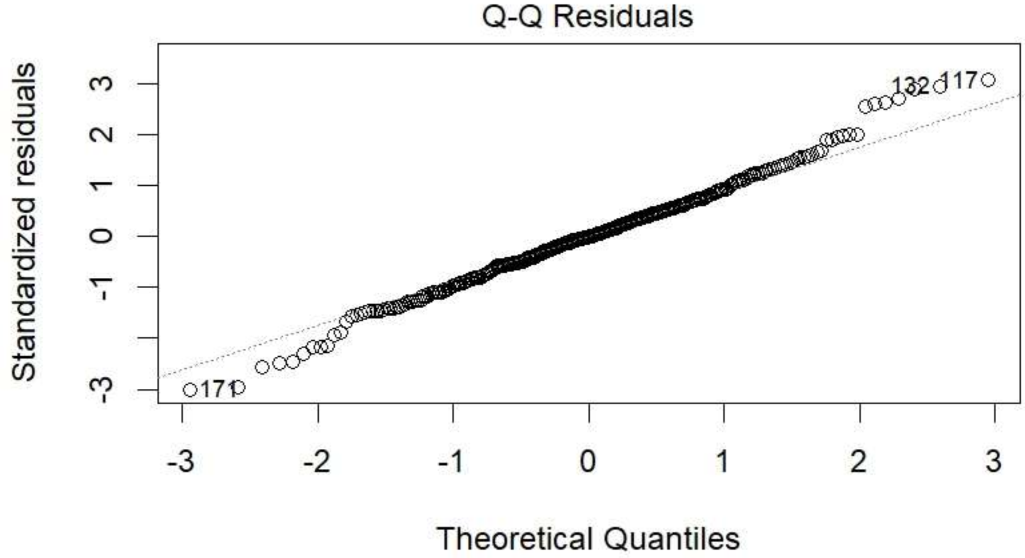


**Figure: 4.6 Q-Q Plot for Residuals**

As Multiple Linear Regression assumptions were met, the Log Model will be approved for our analysis. Predicted data was calculated using a test dataset to validate the model.

Accuracy for prediction was computed using Mean Absolute Percentage Error (MAPE) (Tofallis, 2015).

MAPE = 0.09869431
Accuracy = 1- MAPE = 0.9013057.
The accuracy of prediction is 90% which is considered high.

### 4.4 Final Multiple Regression Model

The approved model is the log model. Its multiple R-squared: is 0.777 and its adjusted R-squa red: is 0.766. This means that this model explains 77% of the variation in the dependent varia ble
(vulnerability index).

As shown in Table 4.9 for the MLR coefficients, significant coefficients will be considered only in the model equation with a P value less or equal to 0.05. It ended with 10 socio-economic indicators which are:
dependency ratio, drinking water, durable goods, electricity, health Insurance, household type, household density, internet capacity, toilet facility, and unemployment. Accordingly, here is the model equation:

**Predicted Vulnerability Index (PVI)** = I+β1(log dependency ratio) + β2(log drinking water)+ β3(log durable goods) + β4(log electricity) + β5(log health Insurance) + β6(log household type) + β7(log household density) + β8(log internet capacity + β9(log toilet facility) + β10(log unemployment)

**PVI** = 25.8205 + (-4.853* log dependency ratio) + (1.6132*log drinking water) + (1.8754* log durable goods) + (1.1065* log electricity) + (-0.8824* log health Insurance) + (0.495* log household type) + (1.7236* log household density) + (3.6096* log internet capacity) + (0.7479* log toilet facility) + (0.9438* log unemployment).

### 4.5 Analysis using ArcGIS Pro

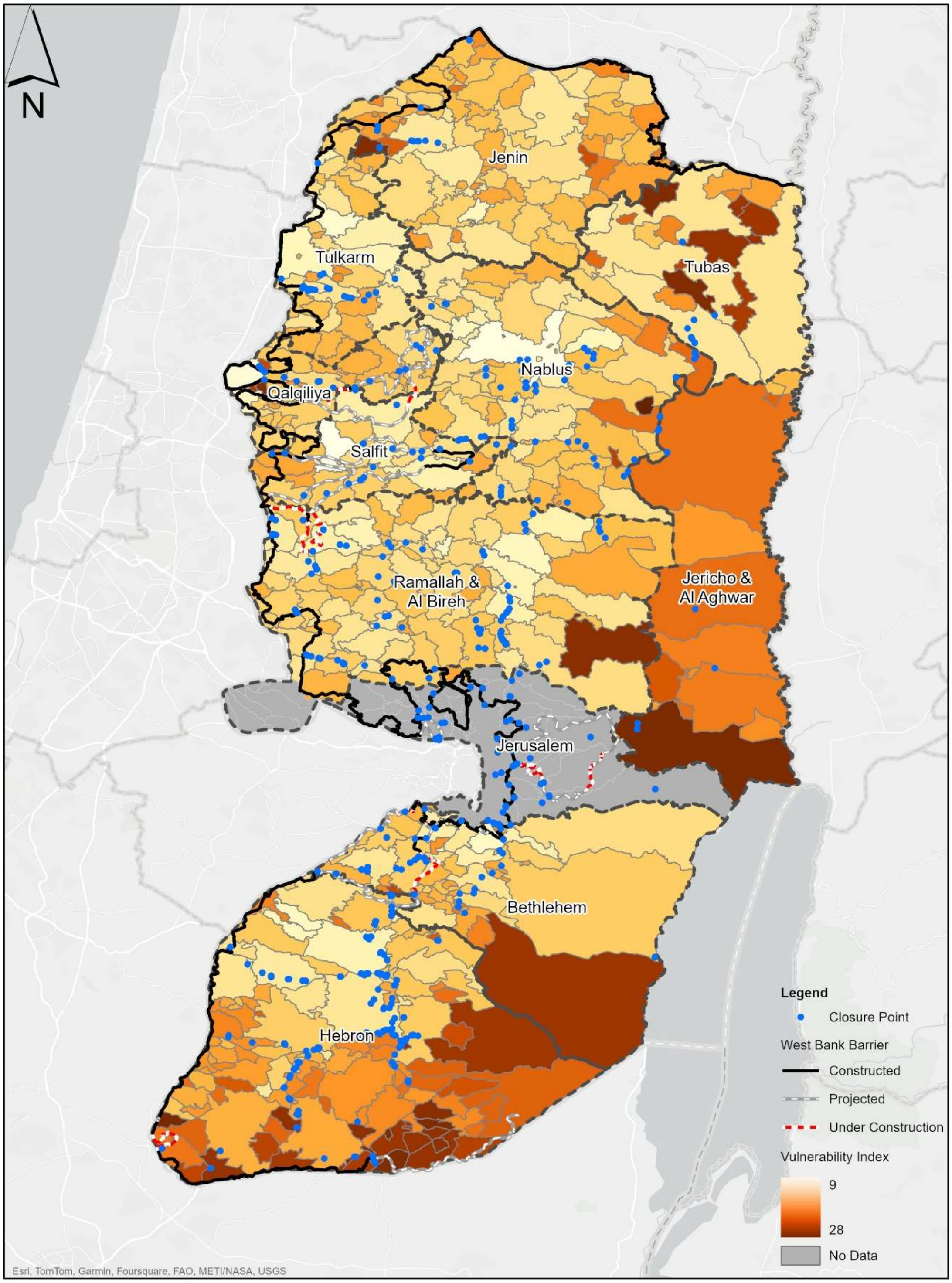
ArcGIS Pro is a geographic information system (GIS) application that is widely used for mapping, spatial analysis, and data management. In this study, it's used to show the PVI for each community. In addition to that, occupation features are located on the map with this index to try to correlate visually with each other. To enhance this spatial analysis, a statistical test was done to check these relationships. Map 4.1 below, shows the vulnerability index with occupation features such as the West Bank Barrier, closures, and settlements.

The vulnerability index ranges from 9-28. It was divided into three ranges, low (9-15.3), medium (15.4-21.5), and high (21.6-28) by implementing a percentile-based approach to set range thresholds involves dividing the data into three percentiles: the 33rd, 66th, and 100th percentiles. The low range spans from the minimum value to the 33rd percentile, the medium range encompasses values between the 33rd and 66th percentiles, and the high range includes values between the 66th and 100th percentiles. S. Rajesha et al. (2018) employed the percentile method to divide the vulnerability index into five ranges.

Table 4.10 shows for each vulnerability range the population, number of communities, and percentage of each range per Governorate.

**Table (4.10) Descriptive Statistics for Vulnerability Index**

| Vulnerability Index | Low (9-15.3) | Medium (15.4-21.5) | High (21.6-28) |
|---|---|---|---|
| **Population** | 1’691’058 | 746’009 | 12’078 |
| **Number of Communities** | 215 | 254 | 52 |
| **Jenin** | 14% | 20% | 6% |
| **Tubas** | 3% | 3% | 15% |
| **Tulkarem** | 9% | 7% | 2% |
| **Qalqiliya** | 7% | 6% | 8% |
| **Salfit** | 5% | 3% | 0% |
| **Nablus** | 17% | 11% | 4% |
| **Ramallah** | 27% | 8% | 2% |
| **Jericho** | 0% | 5% | 4% |
| **Bethlehem** | 9% | 11% | 4% |
| **Hebron** | 9% | 26% | 57% |



**Predicted Vulnerability Index in the West Bank**

0 5.5 11 22 33
Kilometers

Data Sources: PCBS, OCHA 2017

**Map 4.1**

In map 4.1, the vulnerability index score is shown by color. It ranges between 9 – 28. The darker the color is, the more vulnerable is the community.

It can be noticed that one of the most vulnerable areas in the West Bank is south of Hebron governorate. This area is well known as Masafer Yatta. With more investigation, these communities suffer from settler violence and destruction of property.

Another area is east of West Bank, it's called Jordan Valley which also suffers from settler violence and destruction of property. In addition to that, its Area C**.** Chi-square test was done to find out if there are association between the vulnerability index and the occupation features. The test was done for settlements existence, West Bank Barrier, and Area C.
Here are the test results:

1. Vulnerability score of Unemployment Rate VS WBB: As West Bank Barrier was a major occupation feature that prevented freedom of movement and directly affected the unemployment rate. All communities that WBB passed by it were highlighted. The vulnerability score the of Unemployment rate was categorized into three ranges, low (0.0037-0.99), medium (1-1.99), and high (2-3.14). Implementing a percentile-based approach to set range thresholds involves dividing the data into three percentiles: the 33rd, 66th, and 100th percentiles. The low range spans from the minimum value to the 33rd percentile, the medium range encompasses values between the 33rd and 66th percentiles, and the high range includes values between the 66th and 100th percentiles.

   A cross-tabulation table was calculated and used for the Chi-Square test. P value for this occupation feature = 0.000234. The hypothesis for this test is:

   $H_0$: There's no association between WBB and the vulnerability score of Unemployment rate.
   $H_a$: There is an association between WBB and vulnerability score of unemployment rate.

   Cross Tabulation Table:

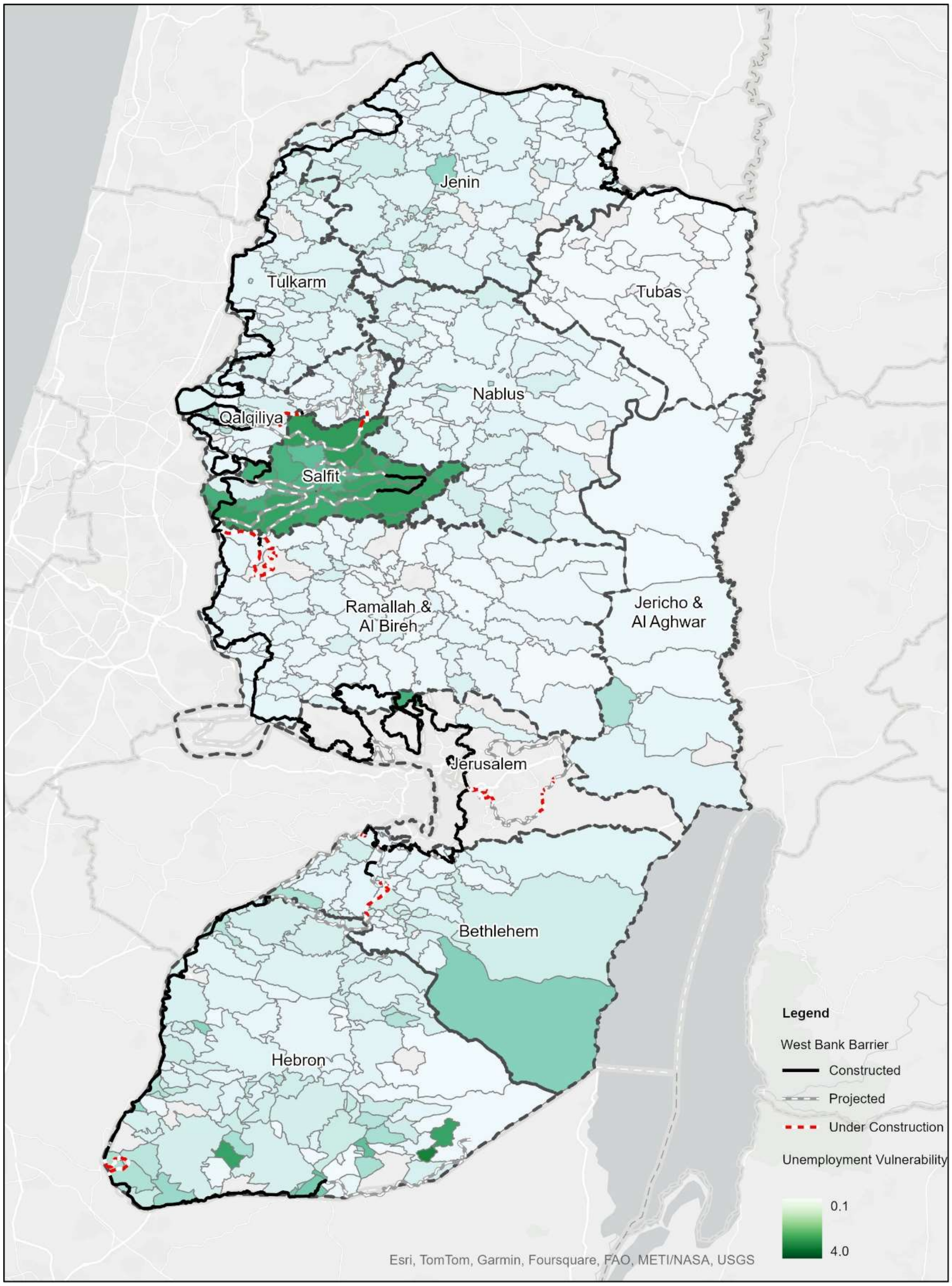
   Unemployment rate VS Existence of WBB

   | | **Low** | **Medium** | **High** |
   |---|---|---|---|
   | **Yes** | 117 | 14 | 16 |
   | **No** | 308 | 30 | 8 |

Pearson's Chi-squared test for unemployment rate

X-squared = 16.72, df = 2, p-value = 0.000234

As the p-value is less than 0.05, then we reject $H_0$. It implies that there is an association between WBB and the vulnerability score of the unemployment rate. Map 4.2 shows the WBB and vulnerability score of the unemployment rate.



Esri, TomTom, Garmin, Foursquare, FAO, METI/NASA, USGS

**Unemployment Rate VS West Bank Barrier**

Data Sources: PCBS, OCHA 2017

**Map 4.2**

2. Settlement Existence: In this regard, all communities that have settlements on their lands were highlighted. The Vulnerability Index was categorized into three ranges, low (9-15.3), medium (15.4-21.5), and high (21.6-28) as explained above.
   A cross-tabulation table was calculated and used for the Chi-Square test.
   P value for this occupation feature = 0.05. The hypothesis for this test is:

   $H_0$: There is no association between settlement existence and vulnerability index.
   $H_a$: There is an association between settlement existence and vulnerability index.

   Cross Tabulation Table:

Vulnerability Index VS Existence of Settlements

| | Low | Medium | High |
|---|---|---|---|
| **Yes** | 102 | 91 | 25 |
| **No** | 113 | 169 | 29 |

Pearson's Chi-squared test

X-squared = 8.1616, df = 2, p-value = 0.01689

As the p-value is less than 0.05, then we reject $H_0$. It implies that there is an association between settlement existence and vulnerability index. Map 4.2 shows the communities with settlement existence and communities without. As shown in the legend, communities outlined with black lines with settlements, and the ones outlined with blue lines are without settlements.

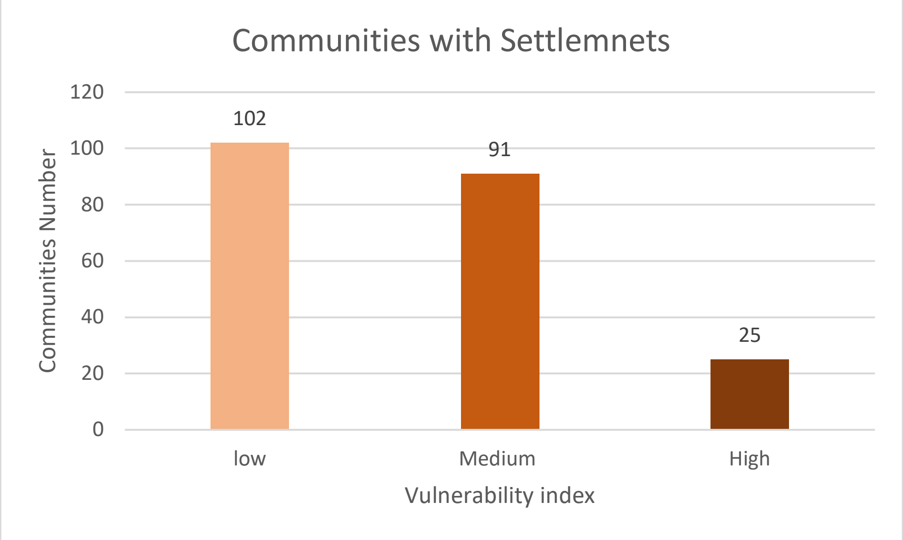
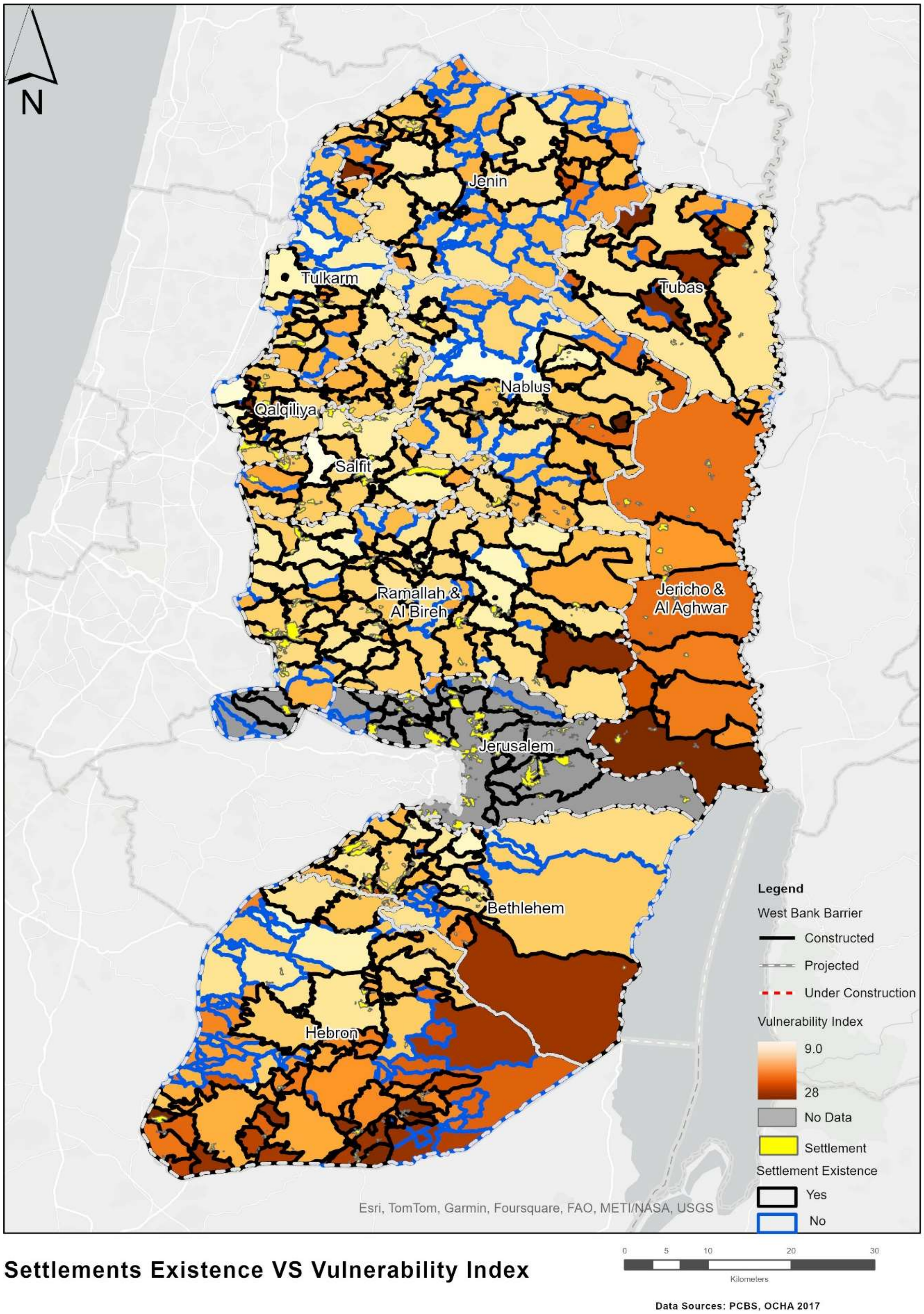


**Figure 4.1 Communities with Settlement Existence**

Figure 4.1 shows that 47% of communities with settlements have a relatively low vulnerability index. 42% of communities with settlements have a medium vulnerability index. 11% of communities with settlements have a high vulnerability index.



Settlements Existence VS Vulnerability Index

Data Sources: PCBS, OCHA 2017

Map 4.3

2. West Bank Barrier: the same procedure was done for the West Bank Barrier (WBB). Communities where WBB passes by were selected. A cross-tabulation table was calculated and used for the Chi-Square test. P value for WBB = 0.15. The hypothesis for this test is:

   $H_0$: There is no association between WBB and vulnerability index.
   $H_a$: There is an association between WBB and vulnerability index.

   Cross Tabulation Table:

Vulnerability Index VS Existence of WBB

| | **Low** | **Medium** | **High** |
|---|---|---|---|
| **Yes** | 63 | 71 | 13 |
| **No** | 145 | 177 | 24 |

Pearson's Chi-squared test

X-squared = 0.6894, df = 2, p-value = 0.7084

As the p-value is more than 0.05, then we don't reject $H_0$. It implies that there is no association between WBB and the vulnerability index.

3. Area C Existence: the same procedure was done for Area C. Communities where com munities are considered under Area C. The Cross tabulation table was calculated and used for the Chi-Square test. P value for Area C = 2.122e-08. The hypothesis for this t est is:

   $H_0$: There is no association between Area C and the vulnerability index.
   $H_a$: There is an association between Area C and the vulnerability index.

   Cross Tabulation Table:

Vulnerability Index VS Existence of Area C

| | **Low** | **Medium** | **High** |
|---|---|---|---|
| **Yes** | 74 | 108 | 27 |
| **No** | 134 | 146 | 4 |

Pearson's Chi-squared test

X-squared = 29.326, df = 2, p-value = 4.284e-07

As the p-value is less than 0.05, then we reject $H_0$. It implies that there is an association between Area C and the vulnerability index.

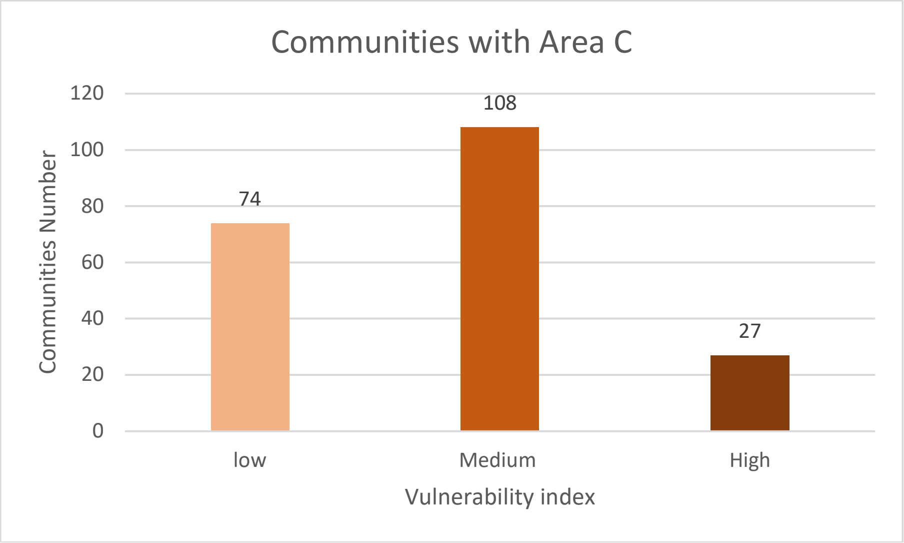


**Figure 4.2 Communities with Area C**

Figure 4.1 shows that 35% of communities with settlements have a relatively low vulnerability index. 52% of communities with settlements have a medium vulnerability index. 13% of communities with settlements have a high vulnerability index.

# Chapter 5: Conclusions and Recommendations

In this chapter, the findings of the study are discussed about the research questions and goal. The primary goal of this study is to create a decision-making tool utilizing socio-economic data from 2017 gathered by the Palestinian Bureau of Statistics (PCBS) to calculate the vulnerability index for the Palestinian communities. Furthermore, the analysis will incorporate occupation features such as the West Bank Barrier, Area C, and settlements with their association with the vulnerability index using statistical tests.

## 5.1 Discussion

The significance of this study lies in its concentration on socio-economic vulnerability within the West Bank, necessitating a donor response tailored to this context. A tool is set to be created for assessing the vulnerability index, acting as a decision-making aid. This tool will aid the government and donors in prioritizing based on needs and conducting data analysis regarding the long-lasting consequences of occupation.

Map 4.1 provides a visual representation of the vulnerability index. It shows that one of the West Bank's most vulnerable is located to the south of the Hebron governorate, commonly referred to as Masafer Yatta. Another area displaying a high vulnerability index is situated to the east of the West Bank, known as the Jordan Valley. Upon closer examination, it becomes evident that these communities suffer from settler violence and destruction of property.

This result aligns with the observations made by (ECHO, 2019), which highlight that the most vulnerable communities are the ones at risk of forced displacement, including Bedouins, and residents in and around Hebron. These communities are the most vulnerable with households experiencing the impact of demolition and confiscation of private property, placing their livelihoods at significant risk.

The first research question investigated the correlation between the unemployment rate and the West Bank barrier. Statically it was tested to be positive. According to the (UN, 2016), access restrictions such as closures and WBB reduced economic activity among the divided Palestinian communities in the region, directly impacting the unemployment rate.

The other two questions investigated the correlation between settlement existence and Area C existence. Statically, both was tested positive. Settlement existence of more severe mobility restrictions, and settler violence, indirectly lead to food insecurity which makes the community more vulnerable (FAO & UN WFP, 2009).
(UN, 2016) mentioned that Area C communities face big obstacles in fundamental rights such as movement, adequate housing, health rights, education, employment, a decent standard of living, and access to justice. The restricted access to water and electricity negatively affects the livelihoods of those relying on agriculture. Due to these facts, residents of Area C are more vulnerable compared to many others.

## 5.2 Conclusions

A multiple linear regression model was used in predicting the vulnerability index using the 18 socio-economic indicators. To meet the multiple linear regression assumptions such as normality, homoscedasticity, and multicollinearity, transformation for data was required. Data was transformed into logarithmic form.

The final logarithmic model was approved, out of 18 indicators, 10 indicators were significant which are: dependency ratio, drinking water, durable goods, electricity, health Insurance, household type, household density, internet capacity, toilet facility, and unemployment.

Final model R-squared: **0.777** and adjusted R-squared: **0.766** which means that the model. explains 77% of the variability of the dependent variable. It also indicates a good fit of the model to the data. Accuracy for prediction = 90%.

Moreover, the PVI was interpreted visually using ArcGIS Pro software. Map 4.1 shows the PVI for each Palestinian community. The darker the color, the more vulnerable is the community.

The correlation between the unemployment rate and the West Bank Barrier was examined. A positive outcome was obtained through a Chi-Square test. Map 4.2 displays the varying levels of unemployment intensity across different communities.

The correlation between the settlement existence of PVI was examined.
A positive outcome was obtained through a Chi-Square test. Visually, Map 4.3 displays the varying levels of PVI VS settlement existence all across communities.

- 39% of communities with settlements have a relatively low vulnerability index.
- 48% of communities with settlements have a medium vulnerability index.
- 13% of communities with settlements have a high vulnerability index.

The correlation between Area C's existence and PVI was examined. A positive outcome was obtained through a Chi-Square test.

- 26% of communities with settlements have a relatively low vulnerability index.
- 50% of communities with settlements have a medium vulnerability index.
- 24% of communities with settlements have a high vulnerability index.

The correlation between the West Bank Barrier and PVI was examined. A negative outcome was obtained through a Chi-Square test. No association between the West Bank Barrier and th e PVI

## 5.3 Recommendations

The study's findings lead to several recommendations for future initiatives linked to its subject matter. Firstly, it suggests utilizing socio-economic data from 2007 (if available) and conducting a similar analysis with 2017 data, facilitating a comparison to discern changes or discrepancies and their potential correlation with occupation features. Secondly, Identify communities with a high vulnerability index and intensify the analysis by exploring deeper into socio-economic indicators. Analyze these indicators separately and correlate the results with additional occupation features, such as physical closures, access to land, seam zone (lands behind WBB), and house demolition. Thirdly, it proposes the inclusion of climate change data in the analysis, given its significant impact on the agricultural sector and factors like water availability, which directly influence community vulnerability. Lastly, it recommends that governments and international entities should employ this tool to enhance the prioritization of aid based on the level of community vulnerability, ultimately resulting in the reduction of this vulnerability.

## References

A. Jaafari, D. Mafi-Gholami, S. Yousef. (2023). A spatiotemporal analysis using expert-weighted indicators for assessing social resilience to natural hazards. *Sustainable Cities and Society*.

Allen, K. (. (2003). Vulnerability reduction and the community-based approach: A Philippines study. Dans M. Pelling, *Natural disaster and development in a globalizing world* (pp. 170–184). ed. M. Pelling, 170–184, New York: Routledge.

Aubrecht, Freire, Neuhold & others. (2012). Introducing a temporal component in spatial vulnerability analysis. Disaster Advances, 5(2), 48-53.

Commission, E. (2020). *Knowledge of Policy*. Récupéré sur https://knowledge4policy.ec.europa.eu/composite-indicators/10-step-guide/step-6-weighting_en#equal-weights

Daoud, J. I. (2017). Multicollinearity and regression analysis. In. IOP Publishing. *Journal of Physics: Conference Series* , (Vol. 949, No. 1, p. 012009).

ECHO, D.-G. f. (2019). HUMANITARIAN IMPLEMENTATION PLAN (HIP).

El-Sughayyar, Ghattas, Hrimat, & others. (2013). Food security and health assessment of vulnerable households in the occupied Palestinian territory: a cross-sectional study. The Lancet, 382, S10.

FAO & UN WFP. (2009). Comprehensive Food Security and Vulnerability.

Gerra et al. (2020). *Socioeconomic status, parental education, school connectedness and individual socio-cultural resources in vulnerability for drug use among students. International journa.* nternational journal of environmental research and public health, 17(4), 1306.

Gujarati, D. N. (2021). *Essentials of Econometrics, 5th edition.* New York: McGraw-Hill.

Kozel, Fallavier, & Badiani. (2008). Risk and vulnerability analysis in world bank analytic work. Soc. Prot. Discuss. Paper. World Bank, 68.

Laukkonen J, B. P.-N. (2009). Combining climate change adaptation and mitigation measures at the local level.

Mansur, A. B. (2016). . An assessment of urban vulnerability in the Amazon Delta and Estuary: a multi-criterion index of flood exposure, socio-economic conditions and infrastructure. . 625–643.

OCHA. (2022). *OCHA'S Strategic Plan 2023-2026.* OCHA's Policy Branch.

PCBS. (2006). *Survey of Israeli Unilateral Measures on the Social, Economic and Environmental Conditions 2006.* Ramallah, Palestine: Palestinian Bureau of Statistics.

PCBS. (2016). Survey on the Impact of Israeli Aggression on Gaza Strip, 2014. Ramallah, Palestine: Palestinian Bureau of Statistics.

Raju, K. V. (2016). Socio-economic and Agricultural Vulnerability Across Districts of Karnataka. Climate Change Challenge (3C) and Social-Economic-Ecological Interface-Building.

S. Rajesha, S. Jain, P. Sharma. (2018). Inherent vulnerability assessment of rural households based on socio-economic indicators using categorical principal component analysis:. *Ecological Indicators*, 93–104.

Smith, E. F. (2014). Socio-economic Vulnerability Assessment of the Burnett-Mary Horticultural Sector. Australia.

Šoltés, V. Š. (2016). Socio-economic analysis of development of regions. Global Journal of Business, Economics and Management: . Current Issues, 6(2), 171-178.

Sorg, Medina, Feldmeyer & others. (2018). Capturing the multifaceted phenomena of socioeconomic vulnerability. *Capturing the multifaceted phenomena of socioeconomic vulnerability.* Natural Hazards.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. . *Journal of the Operational Research Society*, 1352-1362.

UN. (2016). *Leave No One Behind: A Perspective on Vulnerability and Structural Disadvantage in Palestine.* United Nations Country Team.

UNICEF. (2018). *Education and adolescents.* UNICEF.

UNISDR. (2015). United nations international strategy on disaster reduction, sendai framework for disaster risk reduction 2015–2030, Geneva, Switzerland.

Wahyuni, A. T. (2022). Wahyuni, A. T., Rachmawati, R.,Spatial Analysis of Socio-Economic Vulnerability in COVID-19 Handling: Strategies for the Development of Smart Society and Smart Economy. . Information, 13(8), 366.

## Appendix (A) R codes

```
attach(indicators_only)                  # attach the data file#
psych::describe(indicators_only)   # find descriptive stats#
sum(is.na(indicators_only))           #check for missing data#

library(ggplot2)                         # check for outliers#
boxplot(`Dependency Ratio`, main='Dependency Ratio')
boxplot(`Household Size`, main='Household Size')
boxplot(`Educational Attainment`, main='Educational Attainment')
boxplot(`Refugee Status`, main='Refugee Status')
boxplot(`Household Type`, main='Household Type')
boxplot(Unemployment, main='Unempoyment')
boxplot(`Female labour Force`, main='Female Labour Force')
boxplot(Disability, main='Disability')
boxplot(`Tenure of Housing`, main='Tenure of Housing')
boxplot(`Housing Density`, main='Housing Density')
boxplot(`Type of Housing`, main='Type of Housing')
boxplot(`Health Insurance`, main='Health Insurance')
boxplot(Transportation, main='Transportation')
boxplot(`Durable Goods`, main='Durable Goods')
boxplot(`Drinking Water`, main='Drinking Water')
boxplot(Electricity, main='Electricity')
boxplot(`Toilet Facility`, main='Toilet Facility')
boxplot(`Internet Capacity`, main='Internet Capacity')

library(carData)
library(car)
library(lattice)
library(ggplot2)
library(lava)
library(caret)
attach(VulValues)
set.seed(123)                # set data seed
RData<-runif(nrow(VulValues))   # sort data randomly
R<- VulValues [order(RData), ]  # sort data randomly
```

```
Train.Values<-R [1:314, ]    # Split data to train dataset
Test.Values<-R [315:521, ] # Split data to test dataset
MLR1<-lm(`Vulnerability index`~`Vul Dependency Ratio`+`Vul disability`+`Vul
Drinking Water`+`Vul Durable Goods`+`Vul Educational Attainment`+`Vul
Electricity`+`Vul Female out labour Force`+`Vul Health Insurance`+`Vul Household
Size`+`Vul Household Type`+`Vul Housing Density`+`Vul Internet Capacity`+`Vul
Refugee Status`+`Vul Tenure of Housing`+`Vul Toilet Facility`+`Vul
Transportation`+`Vul Type of Housing`+`Vul Unemployment`, data = Train.Values)
# Multiple Linear Regression for 18 indicator using Train data
summary((MLR1))  # Show Model Summary
vif(MLR1)        # Show Variance Inflation Factor for predictors for Model 1
plot(MLR1)       # Plot Model1
shapiro.test(residuals(MLR1)) # Normality test for Model 1
MLR2<-lm(`Vulnerability index`~`Vul Dependency Ratio`+`Vul disability`+`Vul
Drinking Water`+`Vul Durable Goods`+`Vul Educational Attainment`+`Vul
Electricity`+`Vul Female out labour Force`+`Vul Health Insurance`+`Vul Household
Size`+`Vul Household Type`+`Vul Housing Density`+`Vul Internet Capacity`+`Vul
Refugee Status`+`Vul Tenure of Housing`+`Vul Toilet Facility`+`Vul
Transportation`+`Vul Unemployment`, data = Train.Values)  # Multiple Linear
Regression for 17 indicator using Train data
summary((MLR2))  # Show Model Summary
vif(MLR2)        # Show Variance Inflation Factor for predictors for Model 2
plot(MLR2)       # Plot Model2
shapiro.test(residuals(MLR2)) # Normality test for Model 2
MLR3<-lm(`Vulnerability index`~`Vul Dependency Ratio`+`Vul disability`+`Vul
Drinking Water`+`Vul Durable Goods`+`Vul Educational Attainment`+`Vul
Electricity`+`Vul Female out labour Force`+`Vul Health Insurance`+`Vul Household
Size`+`Vul Housing Density`+`Vul Internet Capacity`+`Vul Refugee Status`+`Vul
Tenure of Housing`+`Vul Toilet Facility`+`Vul Transportation`+`Vul
Unemployment`, data = Train.Values)  # Multiple Linear Regression for 16 indicator
using Train data
summary((MLR3))  # Show Model Summary
vif(MLR3)        # Show Variance Inflation Factor for predictors for Model 3
plot(MLR3)       # Plot Model2
shapiro.test(residuals(MLR3)) # Normality test for Model 3
MLR4<-lm(`Log Vulnerability index`~`Vul Dependency Ratio`+`Vul
disability`+`Vul Drinking Water`+`Vul Durable Goods`+`Vul Educational
Attainment`+`Vul Electricity`+`Vul Female out labour Force`+`Vul Health
Insurance`+`Vul Household Size`+`Vul Housing Density`+`Vul Internet
Capacity`+`Vul Refugee Status`+`Vul Tenure of Housing`+`Vul Toilet
Facility`+`Vul Transportation`+`Vul Unemployment`, data = Train.Values)  #
Multiple Linear Regression for 16 indicator using Train data and Log for dependent
summary((MLR4))  # Show Model Summary
vif(MLR4)        # Show Variance Inflation Factor for predictors for Model 4
plot(MLR4)       # Plot Model 4
shapiro.test(residuals(MLR4)) # Normality test for Model 4
attach(VulValues_Log)
set.seed(123)
RDatalog<-runif(nrow(VulValues_Log))    # sort data randomly
RLog<- VulValues_Log [order(RDatalog), ]  # sort data randomly
```

```
Train.LogValues<-RLog [1:314, ]    # Split data to train dataset
Test.LogValues<-RLog [315:521, ] # Split data to test dataset
attach(VulValues_Log)
MLRLOG<-lm(`Vulnerability index`~`Log Vul Dependency Ratio`+ `Log Vul
disability`+`Log Vul Drinking Water`+`Log Vul Durable Goods`+`Log Vul
Educational Attainment`+`Log Vul Electricity`+`Log Vul Female out labour
Force`+`Log Vul Health Insurance`+`Log Vul Household Size`+`Log Vul Household
Type`+`Log Vul Housing Density`+`Log Vul Internet Capacity`+`Log Vul Refugee
Status`+`Log Vul Tenure of Housing`+`Log Vul Toilet Facility`+`Log Vul
Transportation`+`Log Vul Type of Housing`+`Log Vul Unemployment`, data =
Train.LogValues )    # Multiple Linear Regression for 18 indicator using Log Train
data
summary(MLRLOG)      # Show Model Summary
vif(MLRLOG)
MLRLOG1<-lm(`Vulnerability index`~`Log Vul Dependency Ratio`+ `Log Vul
disability`+`Log Vul Drinking Water`+`Log Vul Durable Goods`+`Log Vul
Educational Attainment`+`Log Vul Electricity`+`Log Vul Female out labour
Force`+`Log Vul Health Insurance`+`Log Vul Household Size`+`Log Vul Household
Type`+`Log Vul Housing Density`+`Log Vul Internet Capacity`+`Log Vul Refugee
Status`+`Log Vul Tenure of Housing`+`Log Vul Toilet Facility`+`Log Vul
Transportation`+`Log Vul Unemployment`, data = Train.LogValues )    # Multiple
Linear Regression for 17 indicator using Log Train data
summary(MLRLOG1)
vif(MLRLOG1)
MLRLOG2<-lm(`Vulnerability index`~`Log Vul Dependency Ratio`+ `Log Vul
disability`+`Log Vul Drinking Water`+`Log Vul Durable Goods`+`Log Vul
Educational Attainment`+`Log Vul Electricity`+`Log Vul Health Insurance`+`Log
Vul Household Size`+`Log Vul Household Type`+`Log Vul Housing Density`+`Log
Vul Internet Capacity`+`Log Vul Refugee Status`+`Log Vul Tenure of
Housing`+`Log Vul Toilet Facility`+ `Log Vul Transportation`+`Log Vul
Unemployment`, data = Train.LogValues )    # Multiple Linear Regression for 16
indicator using Log Train data
summary(MLRLOG2)
vif(MLRLOG2)
MLRLOG3<-lm(`Vulnerability index`~`Log Vul Dependency Ratio`+ `Log Vul
disability`+`Log Vul Drinking Water`+`Log Vul Durable Goods`+`Log Vul
Educational Attainment`+`Log Vul Electricity`+`Log Vul Health Insurance`+`Log
Vul Household Size`+`Log Vul Household Type`+`Log Vul Housing Density`+`Log
Vul Internet Capacity`+`Log Vul Refugee Status`+`Log Vul Tenure of
Housing`+`Log Vul Toilet Facility`+`Log Vul Unemployment`, data =
Train.LogValues )    # Multiple Linear Regression for 15 indicator using Log Train
data
summary(MLRLOG3)
vif(MLRLOG3)
shapiro.test(residuals(MLRLOG3))      # Normality test for Model 4
plot(MLRLOG3)
PredictMLRLOG<- predict(MLRLOG3,Test.LogValues)   # Validate model with test
data
PredictMLRLOG                   # Show Predicted data
```

```
Test.LogValues ["Predicted"]<-PredictMLRLOG   # Add predicted column to test data
View(Test.LogValues)                    # show predicted column
library(openxlsx2)
library(dplyr)
library(MLmetrics)
View(Test.LogValues)
attach(Test.LogValues)
Error<-MAPE(Test.LogValues$Predicted,Test.LogValues$`Vulnerability index`)  # calculate error in prediction

Error                         # view Error
Accuracy<-1-Error             # calculate accuracy in prediction
Accuracy                      # view accuracy
# create Threshold for Vulnerability Index
Predicted_Values<- seq(9, 28)
low_threshold <- quantile(Predicted_Values, 0.33)  # 33th percentile
medium_threshold <- quantile(Predicted_Values, 0.66)  # 66th percentile (median)
high_threshold <- quantile(Predicted_Values, 1.00)  # 100th percentile
print(paste("Low threshold:", low_threshold))
print(paste("Medium threshold:", medium_threshold))
print(paste("High threshold:", high_threshold))
# create Threshold for Unemployment rate
attach(Uempl)
Uempl<- seq(0.0037,3.14)
low_threshold <- quantile(Uempl, 0.33)  # 33th percentile
medium_threshold <- quantile(Uempl, 0.66)  # 66th percentile (median)
high_threshold <- quantile(Uempl, 1.00)  # 100th percentile
print(paste("Low threshold:", low_threshold))
print(paste("Medium threshold:", medium_threshold))
print(paste("High threshold:", high_threshold))
# Cross tabulation
Settlement<-matrix(c(102,91,25,113,169,29), byrow = T, nrow = 2)
Settlement
colnames(Settlement)<- c("Low","Medium", "High")
rownames(Settlement)<- c("Yes", "No")
Settlement
# Chi Square test
ChiSquare_Sett<- chisq.test(Settlement)
ChiSquare_Sett
# Cross tabulation
WBB<-matrix(c(63,71,13,145,177,24), byrow = T, nrow = 2)
WBB
colnames(WBB)<- c("Low","Medium", "High")
rownames(WBB)<- c("Yes", "No")
WBB
# Chi Square test
ChiSquare_WBB<-chisq.test(WBB)
ChiSquare_WBB
# Cross tabulation
```

```r
Area_C<-matrix(c(74,108,27,134,146,4), byrow = T, nrow = 2)
Area_C
colnames(Area_C)<- c("Low","Medium", "High")
rownames(Area_C)<- c("Yes", "No")
Area_C
# Chi Square test
ChiSquare_AreaC<-chisq.test(Area_C)
ChiSquare_AreaC
# Cross tabulation
Unemp<-matrix(c(117,14,16,308,30,8), byrow = T, nrow = 2)
Unemp
colnames(Unemp)<- c("Low","Medium", "High")
rownames(Unemp)<- c("Yes", "No")
Unemp
# Chi Square test
ChiSquare_Unemp<-chisq.test(Unemp)
ChiSquare_Unemp
psych::describe(Predicted_Values)
```