ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCES

DEPARTMENT OF STATISTICS

STATISTICAL ANALYSIS OF SPATIAL DISTRIBUTION OF TUBERCULOSIS IN NORTH SHOA ZONE, ETHIOPIA

BY

HABTE TADESSE

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULLFILLEMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS

MAY, 2011

ADDIS ABABA, ETHIOPA

TABLE OF CONTENTS

ACKNOWLEDGMENT

My deep gratitude and appreciation goes to Dr. Butte Gotu, my thesis advisor, for all his invaluable and continuous advice and leading me in the right way with detailed and thoughtful comments.

I am also highly indebted to Ato Abebe Bedada for all his encouragement by providing laptop. I thank others without mentioning their names since they are many and I wouldn't be able to do it in only a few of pages. Thank you all.

ABSTRACT

Tuberculosis (TB) is the major cause of health problems in Ethiopia, accounting for more than thousands of cases and deaths occurring annually. The risks of morbidity and mortality associated with TB are characterized by spatial variations across the country. For instance, Oromia is a region with the largest number of TB cases in Ethiopia (36.80%) and the share of North Shoa Zone is really quite big (FMOH, 2008).

This study examines the spatial patterns of TB in North Shoa Zone and identifies those variables that determine TB clustering. We examine the global and local patterns of TB distribution by using individual morbidity data collected from North Shoa Health Bureau in 2008. Global Moran's I, Geary's C and Moran scatter plot are used in determining distribution of TB. These were used in identifying areas of hot spot for giving strong care in monitoring and to reduce TB distribution.

The values for Global Moran's I shows that the presence of significant TB clustering in North Shoa Zone. And in six woredas significant TB clustering of similar values were observed by using cluster map while only in one woreda a significant TB clustering of dissimilar values was observed. Furthermore, Poisson and negative binomial regression analysis are used in this study. The findings of these models have shown that all explanatory variables (population density, number of health centers and prevalence of HIV cases) are significantly associated with TB case loading.

There is evidence of significant TB clustering in North Shoa Zone. Significant hot spots and cold spots of TB clusters were identified in six woredas. Poisson and negative binomial regression analysis show a decrease in TB case loading with increasing number of health centers. Additionally, it is population density that is highly associated with TB case loading.

ACRONYMS

AIC	Akaike Information Criterion
AICC	Corrected Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
BIC	Bayesian Information Criterion
CSA	Central Statistical Agency
FMOH	Federal Ministry of Health
GDP	Gross Domestic Product
GIS	Geographic Information System
HIV	Human Immune Virus
IAR	Institute of Agricultural Research
LISA	Local Indicators of Spatial Autocorrelation
SSA	Sub Saharan Africa
TB	Tuberculosis
USA	United States of America
WHO	World Health Organization

Addis Ababa University
School of Graduate Studies
College of Natural Sciences
Department of Statistics

Title of Thesis: Statistical Analysis of Spatial Distribution of Tuberculosis in North Shoa Zone, Ethiopia

By
Habte Tadesse Likassa

Approved by the Board of Examiners:

_____	……
Department Head	**Signature**
_____	…….
Examiner	**Signature**
_____	……
Examiner	**Signature**

CHAPTER ONE
INTRODUCTION

1.1 Background of the study

1.1.1 Etiology and Mode of Transmission of Tuberculosis

Tuberculosis is a bacterial disease and often deadly infectious disease caused by micro bacterium Tuberculosis in humans. The mode of transmission of mycobacterium species from person to person is well established. Virtually new infections with mycobacterium Tuberculosis are acquired via airborne transmissions. The sources of infections are persons with Tuberculosis of the lung who have coughing and sneezing. Coughing and sneezing produce air droplets containing bacilli. Persons in the same household, or who are in frequent contact with an infections patient have the greatest risk of being exposed to the bacilli (Murray and Lopez, 1996). TB is a serious problem usually and it attacks the lungs but can also affect other parts of the body. It is spread through air when people who have the disease cough, sneeze or spit. Its two main important characteristics are expressed as: uncontrolled growth of cells in the human body that affects the lungs and the ability of these cells to migrate from the original site and spread out to distant sites.

Tuberculosis disease is widely observed in the world and causes problems on human life by decreasing the economy of countries. For instance, the 1990 World Health Organization (WHO) report on the global burden of diseases indicated that TB is ranked as the 2nd most morbidity causing death in the world, and is expected to continue up to 2020. WHO in 1993 also declared that Tuberculosis is a global emergency in which a total of one-third of the world's populations is infected with TB and more importantly, Tuberculosis remains as a principal infectious cause of mortality worldwide, resulting in approximately 1.8 million deaths occurring annually.

In 2001 WHO estimated that 1.86 billion persons were at the risk of TB, 8.74 million people develop TB and two million die. This means that someone somewhere contracts TB in every four seconds and one person dies within 10 seconds. Thus, unless their distribution is known and a great care is given, TB patient can infect 10-15 people per a day. TB kills more adults than any other infectious disease worldwide. It mainly affects people who are active in the economic growth of a country or economically productive age group who are between (15-45 years), thereby causing large social and economical burden on a country. This in turn, hampers the development of a country at large. As a consequence, knowledge about the distribution of TB plays a crucial role in formulating TB prevention and control program.

TB is still a major cause of death worldwide, but the global epidemic is on the threshold of decline, except SSA. There were an estimated 8.8 million new TB cases in 2004, of which 7.4 million occurred in Asia and Sub-Saharan Africa (FMOH, 2008).

Maher and Raviglione (2005) indicate that approximately two million people die from TB annually and an estimated two billion people are infected with TB (WHO, 2006). It is also estimated that 1.7 million people died of TB in 2007 globally, and there were an estimated 9.27 million new cases of Tuberculosis of which the majority were in Asia and Sub Saharan Africa. It is thought that the rates of new Tuberculosis infections and deaths per capita have probably been falling globally for several years now. However, the total number of new Tuberculosis cases is still rising.

The majority of the morbidity and deaths occurred due to TB in the world is in Africa of which the share of Ethiopia is really quite big. However, recent evidence demonstrates that TB prevalence (distribution) and TB death rates are globally decreasing after having reached a peak. Since 2005, the TB incidence rate is in decline in all six WHO regions (FMOH, 2008).

In Ethiopia, TB had been identified as one of the major public health problems in the last five decades. And it is also a well known disease that is spatially distributed in different regions of our country. The 2007 WHO report indicates that the number of TB cases largely increased in Ethiopia with many clinical episodes and deaths occurring annually. Ethiopia is ranking 4th

among 22 high TB burdened countries and is leading in Africa. Ethiopia had an estimate of 314,267 TB cases, with incidence rate of 378 cases per 100,000 populations in the year 2007 (WHO-2007).

According to the 2008 WHO estimate, the incidence of TB in Ethiopia was 541 per 100,000 populations, and the Federal Ministry of Health (FMOH, 2007) hospital statistical data reports that TB is the leading cause of morbidity, the second cause of death and the third cause of hospital admission in Ethiopia. These estimates are not sufficient because of inadequate TB cases reporting in most endemic countries and lack of national wide TB distribution pattern. Therefore, accurate estimates of TB distribution are needed for further planning, implementation and evaluation of TB control program. Hence, there is need for knowing the distribution of TB to optimize the use of limited resources in high risk areas.

HIV infection has been identified as a major risk factor in developing TB. This is because infection with HIV destroys the immune defense mechanisms of the body and is, therefore, an important risk factor for the development of TB. It is estimated that 50-60% of HIV infected people will develop TB disease in their life time in contrast with HIV negative persons, whose life time risk is only 10% (FMOH, 2008). HIV pandemic presents a massive challenge to the control of TB. The synergy between TB and HIV/AIDS is strong in high HIV prevalence populations. Tuberculosis is a leading cause of morbidity and mortality, and HIV is fuelling the Tuberculosis epidemic in Ethiopia. This unprecedented scale of the epidemic of HIV related TB demands concerted and urgent action.

HIV increases susceptibility to infection with micro-bacterium TB, the risk of progression to TB disease, and the incidence and prevalence of TB. It also increases the likelihood of re-infections and relapses of TB and in general TB transmission intensity and temporal variation in Ethiopia is mainly determined by population density, prevalence of HIV diseases and not proportionally allocated of health facilities. Based on these variables variation used to further clarify TB distribution pattern (FMOH, 2008).

The risks of morbidity and mortality associated with TB particularly in semi-arid and high-land regions vary spatially and temporally. In addition to this, the levels of TB risk and its transmission intensity vary from region to region. For instance, Oromia region is one of the affected regions as compared to other regions in our country. According to FMOH report in 2007 more than one-third (36.82%) of the TB patients in Ethiopia is in Oromia region. The eradication of TB is the greatest public health challenge for this region. Knowledge about the distribution and clustering of the disease is the way to reduce its prevalence.

The basic problem in geographical distribution of TB disease for North Shoa Zone, Oromia region, is identification of areas with exceptionally high prevalence to test and to identify the reasons behind high prevalence of the disease. In other words, the problem is to identify hotspot areas or an elevated cluster for events in each woreda. A geographical epidemiology can be defined as the description of spatial patterns of disease morbidity and mortality, as part of descriptive epidemiological studies, with the aim of formulating hypothesis about the etiology disease. Thus, a spatial pattern indicates whether features or attribute values of the disease form a clustered or random pattern across specified levels of administration (e. g woredas). A fundamental question of interest in this study is to know whether the occurrence of TB is random or clustered in some manner, or if there is some sort of regularity. Another research question is to find out whether the TB risk varies spatially depending on variables such as number of health centers, population density and prevalence of HIV cases.

Disease cluster investigation (identification) in space and or time provides information to public health policy makers. The identification of geographical areas with ongoing disease transmission , using GIS and spatio-temporal statistical analysis, has become indispensible. Spatio-temporal clustering methods are concerned with the identification of greater density of occurrences of a phenomenon in certain places and at certain times. Identification of disease clustering is a major interest of epidemiologists; for an effective disease management it is essential to know when, where and to what degree a disease is present.

TB is the leading cause of morbidity and mortality in Ethiopia, accounting for more than thousands of cases and thousands of deaths occurring annually. The risks of morbidity and

mortality associated with TB are characterized by spatial variation across the country. Consequently, we recognize the spatial variation of TB by means of spatial autocorrelation (FMOH, 2008).

Spatial autocorrelation may be defined as the relationship among values of a single variable that comes from the geographic arrangement of the areas in which these values occur. Spatial autocorrelation measures and analyzes the degree of dependency among observations in a geographic space. It needs measuring a spatial weight matrix that reflects the intensity of the geographic relationship between observations in a neighborhood. Spatial autocorrelation statistics such as global Moran's *I* and Geary's *C* are estimates the overall degree of spatial autocorrelation for a data set. The possibility of spatial heterogeneity suggests that the estimated degree of autocorrelation may vary significantly across geographic space. Local indicators spatial autocorrelation (LISAs) provide disaggregated to the level of spatial analysis units. Global spatial autocorrelation statistics such as the global Moran's *I* and Geary's *C* describe the overall spatial dependence of TB overall the entire region, local spatial autocorrelation statistics mainly the local Moran's *I* identified from the Moran scatter plot (Anselin, 1995) is useful in identifying local patterns or hot spots.

Once the distribution is identified, another resolution enhancement in spatial data that offers an opportunity to model disease because it also demands a corresponding enhancement in identifying factors that affect TB distributions. This is mainly used to identify the relation between TB and explanatory variables. The Poisson and negative binomial regression models are used for this purpose and to model spatial data where the spatial effects are used to account spatial dependence.

1.2 Statement of the problem

TB is a serious health problem in North Shoa Zone affecting highly the socio-economic and health status of the Zone at large. One of the research questions in this study is to know whether the spatial distribution of TB disease is random or clustered in different woredas in the study area and to know the most significant variables that determine TB clustering in the study area**.** Another research question is to find out whether Tuberculosis risk varies spatially within different woredas in the North Shoa Zone under the variables taken into consideration**.** Thus, the following problem areas are considered in this study**:**

- Does the spatial distribution of TB spatially random or clustered?
- How do we include spatial dependence in the model?
- How can we model the spatial data under consideration?
- Does TB risk vary spatially and how does it vary?

1.3 Objectives of the study

The general objective of this study is to analyse the spatial pattern of TB in North Shoa Zone and to identify variables that govern the clustering of TB disease.

The specific objectives include**:**

- To determine the distribution of TB in the study area.
- To fit an appropriate model for data under consideration.
- To identify areas with high risk.

1.4 Significance of the study

The results of this study could provide information to government and other concerned organizations in setting policies, strategies and further investigation for understanding the distribution of TB disease. The results help to understand risk factors that are related to the distribution of TB diseases.

CHAPTER TWO

2. LITERATURE REVIEW

Many authors in various disciplines discussed geographical distribution of diseases as a key element in epidemiologic research, depending on importance given to the description of health events such as patients, place and time. Researchers have been focusing on the relationship between demographic factors and health that extremely determine geographical distribution of diseases. The description of spatial patterns of disease incidence and mortality can be defined as geographical epidemiology. Disease mapping has a long history and it is not surprising that this method of descriptive analysis was first used as an attempt to identify disease patterns or sources of infectious disease and to describe rates or intensity of spread.

Barrette and Finke (1792) apparently used the dot map applied to public health problems in describing the distribution of Tuberculosis risk cases in New York.

Snow **(**1954**)** used a point map of TB diseases to investigate the outbreak of TB epidemic in Latino America. GIS and its associated exploratory spatial data analysis were actively used to identify where disease clusters and analyze spatial distribution of diseases in order to evaluate accessibility to health care facilities.

Thomas (1988) an elevated TB incidence occurred in upper copepod Massachusetts. A large epidemiological study on TB made is to investigate the association between TB and several potential risk factors by using geographical and exploratory spatial methods. Different explanatory variables such as age, HIV and population density were used to see their significant effect on TB by using Poisson regression model. The result shows that positive association is observed between age, population density and TB. However, the result also show that HIV/AIDS epidemic is not significantly influencing TB incidence rate of men aged between 30-49 being the most AIDS afflicted population subgroup.

Kulldorf and Negarwalla (1995) made a review of published studies in Antananarivo. The objective of the study was to analyze the spatial distribution of TB in Antananarivo and investigate risk factors. And they developed a new method for the detection and inference of spatial clusters for a particular disease with exploratory spatial analysis mainly Moran's I and local indicators of spatial autocorrelation statistic, with a clearly defined hypothesis test based on the likelihood ratio test. The average incidence during the study time was 74 per 100,000 populations. The results of this study reveal that spatial distribution of TB in Antananarivo was clustered and in three of the six arrondissements (districts) of the city (192 neighborhoods) spatial clustering of high values was observed. A decrease in clustering was observed with movements towards the southern neighborhood. The change in the risk of TB cluster was linked to population density (overcrowding) and patient care factors.

Beyer and Tatley (1996) made a review of published studies on South Africa. The objective of the study was to determine geographical distribution of TB in the two Western Cape Suburbs with the highest reported incidence of TB. The two areas were covering 2.42 km^2 with total population of 34,294. In the study the geographical distribution of cases was determined by GIS based on national population census data obtained in 1991. The result of the study indicates that 1,835 out of 5,345 dwelling units housed at least one case of TB and in 483 houses three or more cases occurred. This result indicates that cases were distributed unevenly through the community. The reasons for the uneven distribution of TB in the community were uncertain. But, it is noticeable from the Poisson regression model that the correlation between crowding and TB case loading was high. In this study spatial distribution of the positive cases was mapped and nearest neighbor method analysis is used to determine the spatial pattern of TB cases.

Kulldorf (1997) used GIS with different exploratory spatial statistics including spatial filtering and cluster analysis to display the spatial patterns of TB disease. Medical and public health professionals have greatly benefited from the use of spatial information as manifested by GIS maps in order to know the spatial distribution of TB diseases over different places. A GIS map was actively used to identify where diseases are highly observed and to determine the spatial distribution of diseases in order to evaluate accessibility to health care facilities where this disease was highly clustered and for making an effective disease management.

Porter (1999) employed GIS technology to identify areas of high TB transmissions and incidence in USA for the period of 1993-1995. Diggle (1993) and Rowlingson (1994) used an alternative approach for analyzing spatially referenced health data, which avoids entirely the need to aggregate data, in point pattern analysis. The data set had subsequently been reanalyzed using a GIS and point pattern analysis to mathematically represent the geographical distribution of observed data that are recorded for a set of points or specific locations.

Friedrich (1998) made a survey on TB distribution and indicated that a geographical data usually exhibit some amount of spatial dependency, a correlation between the values of neighboring districts. In the study, measures for the strength of spatial dependency and tests for the deviation from randomly distributed values, mainly Moran's *I*, Geary's *C* and local indicators of spatial autocorrelation were used. The results of the study indicate that in most districts TB clustering is observed by using local indicators of spatial autocorrelation.

Several authors have used GIS and exploratory spatial data analysis to describe the distribution of various infectious diseases in Africa but only few have focused on TB disease. For instance, Kunneke (1999) in South Africa and Richardson (2002) used combined molecular and spatial tools to understand the transmissions of specific micro bacterium TB strains. While Munche (2003) in South Africa used GIS and spatial exploratory analysis to identify TB transmission patterns in high incidence area and Poisson regression model is fitted. The results based on GIS maps reveals that the major variation incidence of TB was observed in an urban community in Cape Town. Another result of the study was population density highly associated with TB incidence observed in Cape Town.

Paulo and Santos (2005) conducted a study in the city of Ribeirao Preto located in the state of Sao Paulo, Brazil to determine the spatial distribution of TB for the period of 1998-2002. The study used GIS to identify the homes of individuals with TB. TB cases in the urban area of Ribeirao Preto were found to be distributed unequally, concentrated in the North West Region of the city.

Matthew (2005) made a study in Harris County, Texas. The main purpose of the study was to examine the spatial distribution of TB cases in Harris County during the period of 1995-1998 using GIS software. Spatial analytical techniques mainly Global Moran's *I*, Geary's *C* and Local Indicators of spatial autocorrelation were applied. Determining the distribution of TB was based on the incidence rate and intensity measure after checking the assumption of complete spatial randomness. The result of the study indicates that from nine different areas studied for the presence of TB clustering only two were found to have significant clusters and high incidence rate.

Pui et al. (2006) made a study in Taiwan in order to identify spatial anomalies (hotspots) in disease regions. In the study, spatial autocorrelation methods, namely Global Moran's *I* and local indicators of spatial autocorrelation (Moran scatter plot) were used to describe and map spatial clusters. One result of this study indicated that population density and prevalence of HIV could be compared in efforts to formulate the common spatial risk factors for TB distribution.

Tiwari et al. (2006) made a population based study in India. The main aim of the study was to test a large set of TB cases for the presence of statistically significant geographical clusters. Spatial Global Moran's *I* was used to identify spatial distribution of Tuberculosis for the period 2003-2005. The result obtained from this study suggests that there are statistically significant high rates of spatial and space time clusters of TB mainly in three areas of the district. In the study, purely spatial analysis and space time analysis showed the existence of TB clusters was at the same geographical areas of the district.

Carl **(**2007) conducted a case study in Portugal. The purpose of the study was to determine if there are spatiotemporal TB incidence clusters in the country. The study was based on the notification of TB cases between 2000 and 2004. The result of the study indicates that a significant high incidence rate of space time clusters were observed in three areas of Portugal and a purely temporal cluster was identified covering the whole country, during 2002. In the study, Moran scatter plot was used to identify the presence of spatiotemporal disease clustering. An overdispersed Poisson regression model was used where the number of events in an area is Poisson distributed according to a known underlying population at risk.

Asnakew et al. (2007) analyzed malaria clustering in East Shoa, Ethiopia. In the study global spatial autocorrelation and local spatial autocorrelation were used to identify the patterns of malaria distribution in 543 villages. Statistical spatial analysis of malaria incidence by age, temperature and village through time reveal the presence of significant spatio-temporal variations. Poisson regression analysis was used and the result shows that a decrease in malaria incidence with increasing age. The result of local spatial statistics reveals the presence of malaria clustering or hot spots in most villages. Malaria hot spots were identified by using cluster map that are useful for monitoring and targeting of prevention and control measures against the disease.

Alpharetta (2008) analyzed TB clustering in New York. Since the count data exhibited an overdispersion, an overdispersed Poisson regression and negative binomial regression models were used. In the study , the likelihood ratio test, AIC and BIC were used to compare both models and it was concluded that Negative Binomial Regression model is more appropriate than Poisson regression model in fitting a count data that are over dispersed.

Ismael (2008) made a review of cross sectional survey in Alamata Woreda, Southern Tigray. The objective of the study was to know the distribution of TB in two urban and three rural kebeles of the Woreda. The result of the study indicates that the distribution of TB was high in urban areas than in rural areas. However, the Spearman's correlation coefficient analysis indicates no significant association between TB and any of the demographic variables.

Abera et al. (2009) conducted a study in Oromia Regional State, Ethiopia. The purpose of the study was to know the association between HIV and TB and to determine their distribution in the urban and rural areas in the study area. The data for TB case notification and HIV were obtained from Oromia Regional Health Bureau for 17 Zones and 5 Towns of the region. One result of the study indicates that strong positive association was observed between TB and HIV using Spearman's correlation test. Another result of the study indicates that the incidence of TB and HIV was higher in urban areas or towns as compared to the rural areas.

Awash et al. (2009) made a population based cohort study comprising 8,088 malaria cases in Adama, Ethiopia. The study was mainly designed to describe temporal and spatial clustering of malaria cases and to identify factors associated with malaria clustering. One result of the study indicates the existence of stable temporal and spatial malaria clustering in Adama. Global Moran's *I* was used. Another result of the study indicates that among all factors associated with malaria incidence maximum and minimum daily temperatures were the most possible reason for malaria clustering.

Randremanana (2009) reflects that GIS was used to map over 1000 cases of TB in Madagascar and the result of this study shows that significant clustering of TB occurred in Madagascar.

Touray et al. (2010) conducted a study mainly aimed at describing the distribution (pattern) of TB occurrence in Greater Banjul, Gambia. The result of the study indicates that out of 1,145 TB patients, 84% were permanent residents with 88% living in 37 settlements, and significant high and low rate spatial and space time clusters were identified in two districts. In the study, the spatial scan result shows that in Greater Banjul there were significant TB clusters of high rates of TB, and one area with significant low rates. The study has demonstrated the usefulness of spatial analysis in describing the geographical distribution of TB and concluded that there is an evidence of significant clustering of TB cases in Gambia. After an area where disease is clustered was identified, an overdispersed Poisson regression and negative binomial regression models were fitted. The results show that a significant high rate cluster was identified in one of the densely populated areas of the study region.

Knowing the spatial distribution of TB can be a useful tool for putting in place TB control and prevention programs as well as for generating etiologic hypothesis. Spatial autocorrelation is useful for cluster mapping of regional health care problems. Cluster mapping helps to clarify issues such as the spatial aspects of both internal and external correlations of leading health care events which in turn help planners to assess spatial risk factors and identify the most plausible types of health care policies for planning and implementation of health care services.

CHAPTER THREE
DATA AND METHODOLOGY

3.1. Study Area

North Shoa Zone is located between 2,738 and 2,782 meters above sea level at 9^0N, 38^0 E and had an estimated population of 1.26 million in 2007. Since 2008, the zone is administratively organized into 13 woredas and has a density of 3050 people per square kilometer. This makes it the most densely populated zone of the region (CSA, 2007).

North Shoa Zone is one of the 17 zones in the Oromia National Regional State and the main city, Fiche, is located only 130km away from Addis Ababa. The 13 woredas included in this zone are Abicho Nya'a, Aleltu, Dagem, Debre Libanos, Dera, Girar Jarso, Hindhibu Abote, Jida, Kembibit, Kuyu, Wera Jarso, Wuchale, and Yaya Gulalle. The Ethiopian modern capital Addis Ababa is located at the center of this zone.

3.2. Data and Variables of the study: Spatial data may consist of point data such as the locations of events or area data such as locations of road ways. Point data are usually irregularly spaced, but are sometimes aggregated to a regularly spaced grid for convenience. Each item of health data including population at risk because of diseases, environmental exposures, mortality and morbidity are usually connected with spatial point data. Environmental exposures may also be represented as spatially continuous data or as data points at monitoring locations. The data for this study are discrete secondary data or count data obtained from North Shoa Health Bureau documented in 2008 except data on population size. The population size for each woreda of this zone is obtained from a publication of Central Statistical Agency which is documented in 2007.

The independent variables in this study are population density, number of health centers, and prevalence of HIV cases. The dependent variable considered in this study is the number of all forms of TB cases in each woreda (TB case loading).

3.3. Methodology

There appear to be many problems in spatial model where spatial point pattern analysis and modeling are widely applied. The objective of the analysis could to determine if the point pattern of an event is either random or clustered. In order to know the distribution of rare events, exploratory spatial data analysis (Moran's *I*, Geary's *C* and Local Indicators of spatial autocorrelation mainly Moran scatter plot) and Poisson regression model are commonly used. To analyze spatial point pattern data relevant software such as GeoDa, SAS (Statistical Analysis Software) and SPSS program can be used for practical modeling and analysis. Spatial statistics consist primarily of three main components and each of their descriptions is given below.

Geostatistics is a spatial process indexed over a continuous space. Geostatistics began with mining type applications and the prefix "geo" indicates that geostatistics dealt initially with statistics pertaining to the earth. Geostatistics has applications in climatology, environmental monitoring, and geology. One of the most common geostatistical applications in the mining context was kriging (prediction) which deals with predicting the ore grade in a mining area from spatially observed samples. Geostatistical applications have expanded original mining applications to include modeling soil properties, ground water studies, rainfall precipitation, public health, etc (Pfeiffer, 2008).

Lattice Data are spatial data indexed over a lattice data. Lattice type data provide the closest analogue to time series data. In time series data sets, observations are typically obtained at equally spaced time points. For lattice data, spatial data is obtained over a regularly spaced set of points (irregular lattice type data is a possibility as well). Lattice data often come in the form of pixels, which are small rectangular shaped regions, often obtained by remote sensing from satellites or aircraft. Lattice type data for spatial statistics is very similar to the lattice type data one obtains from medical imagining studies, such as PET scans which yield images of the brain by way of pixels or vowels. (Pfeiffer, 2008).

Point Patterns pertains to the location of events in the area of interest (Porto, 1996). Consider a region D and in this region we are interested in locations of certain "events." We can understand whether events of interest are occurring randomly throughout the region, or if the events tend to cluster together. Data analysis of point patterns corresponds to studies where the interest lies in where events of interest occur. A fundamental question of interest in this context is whether or not the points of interest are occurring at random, or do the points cluster in some manner, or perhaps the points of interest are occurring with some sort of regularity. Determining the presence of clustering in a spatial data set can be very difficult since points formed from completely random processes can appear clustered. Spatial point patterns consisting of n events will be compared to completely random point processes. A complete spatial randomness (also known as a homogeneous Poisson process) is one where, conditional on n events (i.e. points in the region), the events are independently and uniformly distributed over the region. Clusters are commonly identified from point pattern data. The objective of examining point patterns is to recognize when events are systematically organized or structured compared with events distributed at random. In general, modeling is preceded by a phase of exploratory spatial analysis, associated to the visual presentation of the data in the form of graphs, maps and identification of spatial dependency patterns in the phenomena under study (John, 1968). However, point data can be lattice or area data when the points are aggregated in the form of area. Point patterns comprise a set of locations of events, usually indexed by geographic coordinates, in a defined study region (Gatrell, 1996). In this case the locations of the events themselves are the phenomena of interest, and the investigation focuses on whether the pattern is exhibiting complete spatial randomness, clustering or regularity (Gatrell, 1996). To determine spatial distribution of events statistical methods and exploratory spatial data analysis have been developed to test whether spatial data are random or not. In case of point pattern analysis the objective of interest is the very spatial location of the events under study.

Generally, spatial data are distinguished by observations that are identified at spatial locations $s_1, s_2, s_3, \dots, s_n$ where s_i coordinates in the plane or polygon are in two dimensions or in three dimensions. Consider spatial data in a region D. The region D is a finite or countable collection of sites where observations are made. The collection of points in D is known as lattice .The spatial sites in the region D are typically identified using its latitude or longitude (Cressie, 1993).

Spatial Dependence

A basic property of spatially located data in a set of values (x_i) is likely to be related over space. Many authors in various disciplines discuss presence of dependence among observations related on diseases in space. Some as cited in Cliff and Ord (1981) are summarized as follow.

Tobler (1970) has referred to "The first law of geography**:** everything is related to everything else but near things are more related than distant things". Stephan (1934) gives the following remark: "data of geographic units are tied together like bunches of grapes, not separated like balls in an urn." These ideas explain why spatial dependence has to be an issue in determining the distribution of TB disease.

Spatial dependence is a key concept in understanding and analyzing a spatial phenomenon. Pattern is that characteristic of the spatial arrangement of objects given by their spacing in relation to each other. It should not be confused with the idea of dispersion, which is relative to some defining area, or with density, which is the average number of objects in a given area. Patterns might consist of clusters of points, a more regular than random arrangement, and trends across real and statistical surfaces and so on. It is believed that spatial dependence is present in every direction and gets weaker the more the dispersion in the data localization increases (Cressie, 1993). Spatial dependence indicates that near places are more likely to be related than distant ones and usually most geographical patterns of interest involve groupings of similar values in clusters.

Spatial autocorrelation analysis is a technique used to detect disease patterns and measures the extent to which the occurrence of an event in areal unit contains or makes more probable to the occurrence of an event in neighboring areal unit. It is defined as the relation among values of a single variable that is attributable to the geographic arrangement of areal units on a map. Spatial autocorrelation is a measure of interdependence between values of a variable at different geographic locations and can be used to identify the degree of clustering (Goodchild, 1987).

Spatial autocorrelation uses a measure known as spatial autocorrelation coefficient to measure and test how clustered or dispersed points are in space with respect to their attribute values. In addition to this, spatial autocorrelation (as a concept that is applied to detect the patterns of points) of a set of points is concerned with degree to which points or things happening at these points are similar to other points or phenomena happening there. If significantly positive spatial autocorrelation exists in a point distribution, points with similar characteristics tend to be near each other. If spatial autocorrelation is weak or non-existent, adjacent points in a distribution tend to have different characteristics. The most widely used measures for the proximity of locations and the similarity of the characteristics of these locations are Moran's I and Geary's C statistics. These statistics mainly measure the strength of spatial autocorrelation among neighboring areal units and are used for testing the assumption of spatial independence or randomness.

3.4. Method of testing spatial randomness

Testing for complete spatial randomness is the first step in the analysis of spatial point pattern data. Basically, the main question here is: are locations randomly distributed through the study area or do the locations indicate some structure? There are several methods and algorithms that are used to answer the scientific question of spatial randomness or clustering of cases. However, quadrant count and nearest neighbor methods are commonly used to test the spatial randomness or clustering of events.

3.4.1. Quadrant Count Method

The basic idea of quadrant method is to divide the region D into subsets often rectangular shape (but other shapes are used as well), and then counting the number of events in each of the subsets. The use of quadrant counts can be used to access whether there is any spatial pattern in the data. If clustering is present in the data, then one would expect quadrants with higher counts to be located near each other. On the other hand, if the quadrant counts are spread out over the region, then there is evidence of uniformity. The quadrant count method can be described simply as partitioning the data set into n equal sized sub regions; these sub regions are called quadrants.

In each quadrant the number of events that occur will be counted and it is the distribution of quadrant counts that will serve as an indicator of pattern. The choice of the quadrant size can greatly affect the analysis, where large quadrants produce a coarse description of the pattern. If the quadrant size is too small then many quadrants may contain only one event or they might not contain any events at all. Usually, the following rule of thumb is used: the area of a square is twice the expected frequency of points in a random distribution (i.e., $2\frac{Area}{n}$), where n is the number of points in the sample size. After partitioning the data set into quadrants, the frequency distribution of the number of points per quadrant will be constructed.

In this study since the shape of the study area cannot be considered as a regular lattice, using exploratory spatial analysis or nearest neighbor method is preferred to quadrant count method.

3.4.2. Nearest neighbor method

This method is commonly applied for irregular areal data which are divided into different polygons (woredas in this study). Identification of polygons which are nearest to each other is a primary concern in exploratory spatial data analysis. In this approach, the nearest neighbor distance method is used to define spatial weight matrix that helps to develop the statistical method used in testing randomness.

Nearest neighbor method gives a flexible weight matrix that represents spatial dependence based on a decay relationship and the number of neighbors. The measurements of areal units that are nearer to each other tend to be similar. When the measurements are independent, then no spatial pattern is expected. The first stage to implement a spatial pattern analysis is the construction and estimation of the weight matrix, given the spatial arrangement of the observations (Anselin and Hudak, 1998).

3.4.2.1. Weight Matrix

A general spatial weight matrix can be defined as a symmetric binary contiguity matrix, which can be generated from topological information based on either adjacency or distance criteria. A fundamental characteristic distinguishing spatial data from time series data is the spatial arrangement of the observations. The spatial linkages or proximity of the observations are measured by defining a spatial weight matrix, denoted by W_{nxn}. The spatial weight matrix represents the strength of the potential interaction between locations. However, it has to be noted that the determination of the proper specification for the elements of a spatial weight matrix is one of the difficult and controversial methodological issues in spatial data analysis (Odland, 1987).

There are two methods that are used in computing spatial weight matrix, namely the Euclidean distance method and the proximity method. The most common method is to consider two or more regions as neighbors if they share a common border or vertex. The proximity or adjacent matrix W, also referred to as the weighting matrix, provides the mechanism for introducing spatial structure in spatial data.

According to the adjacency criteria, the element of spatial weight matrix is 1 if location i is adjacent to location j and 0, otherwise. These will result in $n \times n$ matrix with zeros and ones if the study region has n sites to be investigated. Spatial contiguity for polygons is the property of sharing a common boundary or vertex. Contiguity analysis is an important method for assessing unusual features in the connectivity distribution. The administrative woredas considered in this study are highly irregular in both shape and size. The following examples show how weight matrix is constructed from a regular/irregular study area. Neighborhood relations are defined as either Rooks case, Bishop's case, or Queen's (King's) case.

a. **Rook's case** considers contiguity is by a neighborhood of four locations adjacent to each cell (locations which share a common border are considered as neighbors). As illustrated here on Figure 1 below.

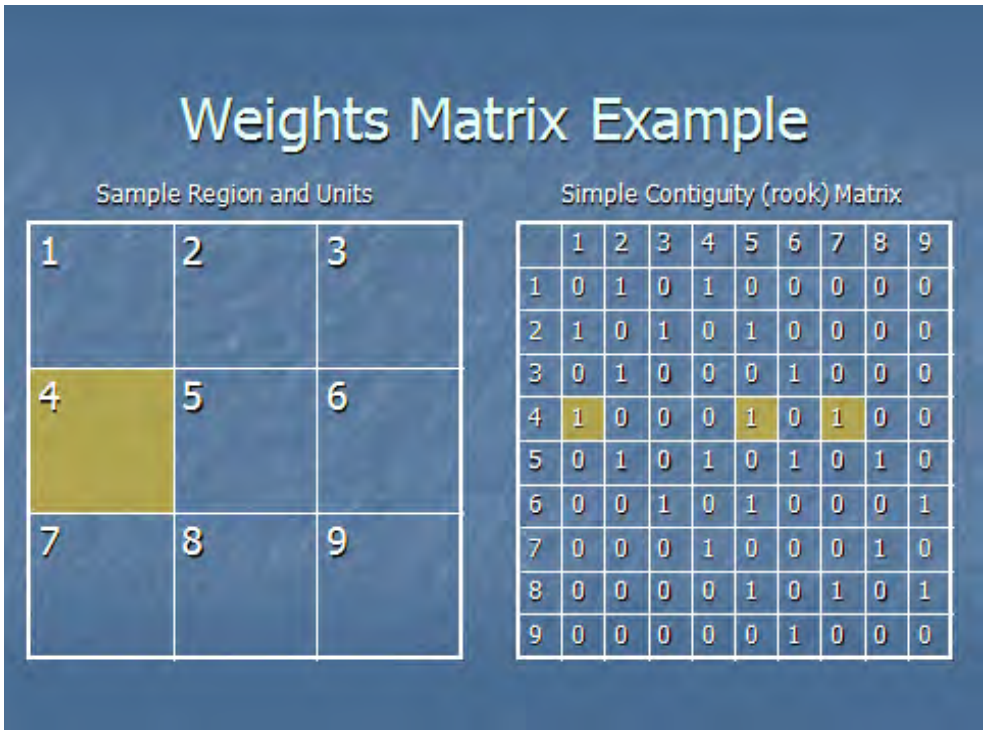


Figure 1. Rook's case

b. Bishop's case only considers the diagonals (common edge) to define a neighborhood as shown in Figure 2 below.

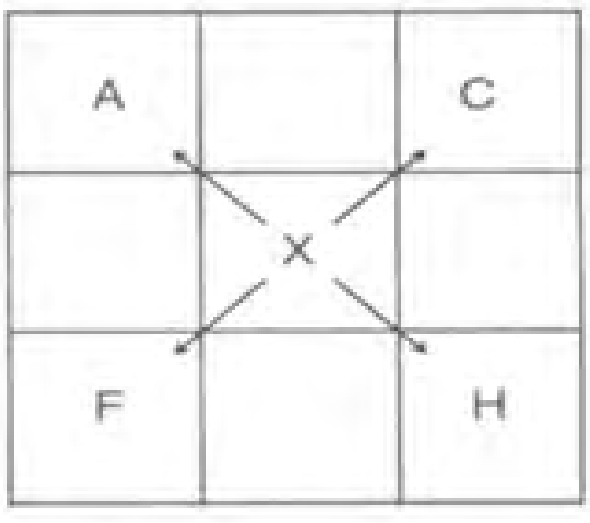


Figure 2. Bishop's case

c. **Queen's case** considers all neighborhoods (common boundary and or edge) to define neighborhood. This is the most commonly used method. Figure 3 below shows an example on how to construct spatial weight matrix based on the Queen's method.

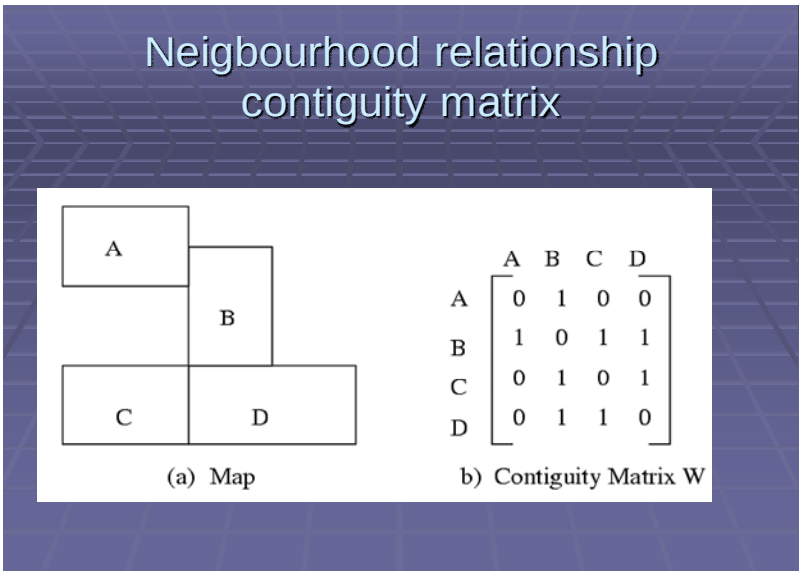


Figure 3. Queen's case

A general cross product used in computing spatial autocorrelation for observations on a variable y is given by:-

$$\gamma = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y}) \qquad [1]$$

where w_{ij} is an element of spatial weight matrix W. There are a number of measures of spatial autocorrelation that are derived from the cross product statistic defined in Equation 1: Moran's *I*, Geary's *C* and local indicators of spatial autocorrelation are the common indices used in computing spatial autocorrelation.

Since the analytical results may be sensitive to the specifications of spatial weight matrix, different spatial weight matrices are applied for different purposes of studies. There is no

universal type of constructing a spatial weight matrix that can be used in spatial data analysis. Constructing spatial weight matrix is obligatory to account for spatial dependence between different polygons and to know the distribution of events using exploratory spatial data analysis. In this study, Global Moran's *I*, Geary's *C* and Local indicators of spatial autocorrelation **(**LISA**)** are used to test for significance of spatial clustering.

3.4.2.2. Global Measures of Spatial Clustering

Moran's *I* Global measures summarize spatial association with respect to the whole region. Spatial autocorrelation index measures spatial association in the data considering simultaneously both locational and attribute information. One of the Global measures of spatial autocorrelation is the well known Moran's *I* given by:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad [2]$$

where n is the number of polygons (woredas in this study), w_{ij} is the element in the spatial weight matrix corresponding to the observation pair i,j or w_{ij} is 1 if the geographic areas associated to y_i and y_j are neighbours and 0 otherwise, and y_i and y_j are observations for areas i and j, with mean $\bar{y}$. The index *I* take on large values if there is a high correlation between neighboring values of the spatial variable i.e. if either large values or small ones (or both) are spatially clustered. The above index has a positive value in case of positive spatial autocorrelation, i.e. when the pairs of deviations from the mean for contiguous locations having the same sign are prevalent. In contrast, when the pairs of deviations from the mean have prevalently opposite sign the index has a negative value, therefore showing negative spatial autocorrelation. However, Moran's *I* indicates a departure from independent observations but doesn't tell where this departure occurs nor even whether large or small values or both are affected since it is applied globally. The observed value of *I* can be compared to its distribution under the null hypothesis of no spatial autocorrelation or no clustering i.e. when the values of y_i are independent of the values $y_j (i \neq j)$ at neighboring locations .This is equivalent to say that

under the reference null distribution, data are randomly distributed over locations. Therefore, inference can be based on the standardized version of *I*, namely

$$Z(I) = \frac{I - E(I)}{\sqrt{Var(I)}} \qquad [3]$$

$$\text{with,} \quad E(I) = \frac{-1}{n-1}\ ,\ Var(I) = \frac{n^2(n-1)s_1 - n(n-1)s_2 - 2s_0{}^2}{(n+1)(n-1)s_0{}^2}$$

$$s_0 = \sum_{i \neq j}^{n} w_{ij}\ ,\ s_1 = \sum_{i \neq j}^{n} \left(w_{ij} - w_{ji}\right)^2\ ,\ s_2 = \sum_{k=1}^{n} \left(\sum_{j=1}^{n} w_{kj} + \sum_{i=1}^{n} w_{ik} \right)^2$$

where k represents districts (woredas).

Interpretation: We can use Moran index for identification of spatial distribution as dispersion, random or cluster patterns. Indices close to zero indicate the presence of random pattern. Indices close to +1 indicate a tendency toward clustering. Besides the fact that Moran's *I* takes the usual form of autocorrelation, its distribution is well studied so that it can be used for testing the significance of spatial autocorrelation in neighboring plots or counties in a study area (Prince, 2010).

In this study, the global Moran's *I* test statistic will be used to test the null hypothesis, H_0 of no significant clustering of TB incidence in the entire study region. The mean found by Moran's *I* coefficient analysis is utilized to identify spatial cluster patterns. A calculated value of Moran's *I* is compared with the tabulated value in order to identify the nature of the distribution of TB. A statistically significant estimate of *I* indicates that neighboring woredas have a similar prevalence rate and that the cases are likely to cluster at the woreda level (Paez, 2004).

Geary's *C*

Geary's *C* interactions are not the cross product of the deviations from the mean, but the deviations in intensities of each observation location with one another**.**

Geary's *C* is given by:-

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(y_i - y_j\right)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left\{(y_i - \bar{y})^2\right\}} \qquad [4]$$

where the notation is the same as in Equation 2. Usually the values of *C* range between 0 and 2. Values of *C* between 1 and 2 indicate presence of negative spatial autocorrelation while values between 0 and 1 indicate presence of positive spatial autocorrelation. Moran's *I* gives a more global indicator, whereas the Geary's coefficient is more sensitive to differences in small neighborhoods**.**

Testing the significance is done by using the standardized version of *C*, namely

$$\text{Z(C)} = \frac{C - E(C)}{\sqrt{Var(C)}} \qquad [5]$$

with $E(C) = 1,\ Var(C) = \frac{\left((2s_1 + s_2)(n-1) - 4{s_0}^2\right)}{2(n+1)s_0}$

where the notation are the same as [3] and $E(C)$, $Var(C)$ are the expectation and variance of *C* coefficients respectively.

Interpretation: The interpretation of Geary's C is analogous to that of Moran's I. The only difference is that when C lies in (1, 2) indicates the presence of negative spatial autocorrelation (TB clustering of dissimilar values), whereas (0, 1) indicates positive spatial autocorrelation

representing the presence of TB clustering of similar values. Smaller *p*-values correspond to stronger autocorrelation for both *I* and *C* statistics. However, in case of Geary's *C*, the *p*-value does not tell us whether the autocorrelation is positive or negative. Based on the preceding remarks, we have positive spatial autocorrelation when $Z_I > 0$ or $Z_C < 0$ and we have negative autocorrelation when $Z_I < 0$ or $Z_C > 0$.

Normality**:** The $y_i's$ (values of a dependent variable) are assumed to be observations on *n* independent drawings from a normal population or populations. The values for the districts (woredas in this study) are derived from a common distribution function, for instance a normal distribution function with a common expectation and variance. It is well known that Moran's *I* and Geary's C are asymptotically normally distributed under the normality assumption for fairly general conditions (Cliff and Ord, 1981). However, if the assumption of Moran's *I* (normality) is doubt we can use the Monte Carlo simulation. Simulate Moran's *I* *n* times under the assumption of no spatial pattern (random) and calculate Moran's *I* then compare actual value of Moran's *I* with the p-value.

3.4.2.3. Local Indicators of Spatial Autocorrelation

While the strength of Moran's I lies in its simplicity, its major limitations is that it tends to average local variations in the strength of spatial autocorrelation. This has encouraged researchers to develop local indices of spatial association. This category of tools examines the local level of spatial autocorrelation in order to identify areas where values of the variable are both extreme and geographically homogeneous. This approach is useful when, in addition to global trends in the entire sample of observations, there exist also pockets of localities exhibiting homogeneous values that do not follow global trend. This leads to identification of hot spots, cold spots and clustering of dissimilar values. Anselin (1995) defines a LISA (local indicators of spatial autocorrelation) statistic satisfying the following two conditions:

- The LISA for each observation measures the extent of sign, when positive spatial clustering of similar values around the observation and when negative spatial clustering of dissimilar values around observations.

- The sum of LISAs for all observation is proportional to a corresponding global indicator of spatial autocorrelation.

The local value of a LISA is computed as:

$$I_i = \frac{\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{(z_i - \bar{z})^2} \qquad [6]$$

From the proportionality condition we get**:**

$$\sum_{i=1}^{n} I_i = \gamma I \qquad \textbf{[7]}$$

where I_i is the LISA statistic for each observation, γ is a scale factor and *I,* is a corresponding global spatial autocorrelation measure. The sum of the LISAs is proportional to a global analog up to a scaling factor. These specific configurations can be first identified from a Moran scatter plot showing observed values against the averaged value of their neighbors. Once a significance level is set, values can also be plotted on a map to display the specific locations of hot spots and potential outliers.

The Moran scatter plot is a useful visual tool for exploratory spatial analysis because it enables us to assess how similar an observed value is to its neighboring observations. Its horizontal axis is based on the values of the observations and is also known as the response axis. The vertical Y axis is based on the weighted average of the corresponding observation on the horizontal X axis**.** The Moran scatter plot provides a visual representation of spatial association (dependence) in the neighborhood around each observation. Depending on their position in the plot, the Moran scatter plot data points express the level of spatial association of each observation with its neighboring ones.

The Moran scatter plot can be divided into four quadrants as it is indicated in Figure 4 below: the top right and the bottom left quadrants contain observations showing positive spatial autocorrelation respectively with high-high and low-low data values indicating presence

of clusters. The top left quadrant contains low values in a neighborhood of high values (low high), while the bottom right quadrant contains high values in a neighborhood of low values (high low). In both cases they are showing values of dissimilar clustering (Anselin, 1996).

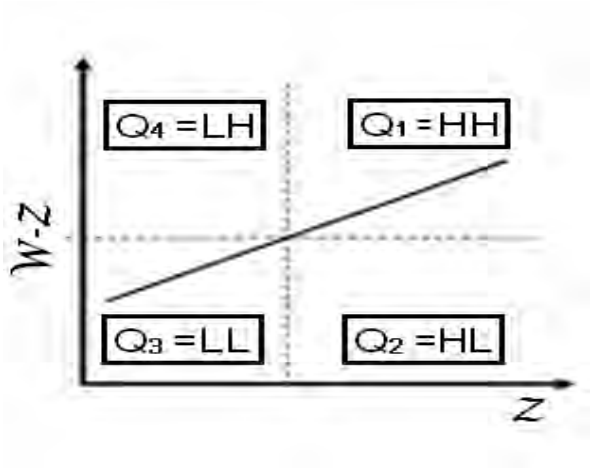


Figure 4. **Moran scatter plot**

3.5. Poisson Regression Model for spatial data

In case where the distribution of TB is clustered, fitting standard Poisson regression model that assumes the variances of TB distribution equal with the mean (randomness assumption) does not hold true because this assumption is implausible in many applications, as events often occur in clusters and data used in count dependent variable has larger variance than mean. In such cases fitting overdispersed Poisson regression and negative binomial regression models are more relevant**.** It is known that Poisson regression model is appropriate and is useful when the outcome is a count type with large counts of rare events (McCullagh and Nelder, 1989).

In spatial model observed counts for a set of areas with a known neighborhood's structure are considered. Thus, in the model specification the observed count is affected by the area where the count is taken and the neighboring areas. For example, the count of people with a particular disease within a set of areas $i = 1,2,3,\dots,n$ can be modeled as Poisson where the random effects can be introduced to account for spatial dependence which is not explained by the observed data. For spatially indexed data, spatial random effects associated with each area or site may be used, allowing for the modeling of an underlying spatial dependence structure.

In a Poisson regression model, observed counts y_i are assumed to have a Poisson distribution, with expected values depending on k predictor variables, $\mathbf{x}= (\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \dots, \mathbf{x_k})^{\mathbf{T}}$. The Poisson regression model aims at modeling a case count dependent variable Y, which follows a Poisson distribution with a parameter μ_i. The probability that the number of cases takes the value on the entity is y_i, and is given by:-

$$f(Y = y_i) = \frac{e^{-\mu_1}\mu_1^{\,y_i}}{y_i!}, y_i = 0,1,2,3.... \qquad [8]$$

For a Poisson distribution the variance is equal to the mean but this assumption is implausible in case where diseases are clustered. The systematic portion of the model involves the explanatory variables $\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \dots, \boldsymbol{x_k}$. Suppose that we want to let the mean μ_i and the variance depends on a vector of explanatory variables x_i. Then, we can obtain a simple linear model of the form:

$$\mu_i = x_i'\beta$$

The above model has a disadvantage that the linear predictor on the right hand side can assume any real value; whereas the Poisson mean μ_i which represents an expected count has to be non-negative. A straight forward solution to this problem is to model the logarithm of the mean using a linear model (Czado, 2008).

Thus, the most common formulation of this model is the log linear model which can be written as:

$$Log\mu_i = x'_i\,\beta \qquad [9]$$

which implies that μ_i is the exponential function of independent variables and can be expressed as:

$$\mu_i = e^{\beta_0+\beta_1x_1+\beta_2x_2+\dots+\beta_kx_k} \qquad [10]$$

The Poisson distribution depends on a single parameter μ_i. Although there is no theoretical upper bound for Poisson distribution, in practice these probabilities could be considered as negligible when y_i is very large. Thus, one can determine the magnitude of y_i based on the values of μ_i (Bauer et al, 2007).

The major assumption of the Poisson regression model is

$$E(y_i \mid x_i) = \mu_i = e^{x_i'\beta} = Var(y_i \mid x_i) \qquad [11]$$

This assumption is based on the premise that successive events occur independently and at the same rate, which is implausible in many applications especially when events occur in clusters. In this model, the regression coefficient $\boldsymbol{\beta_i}$ represents the expected change in the log of the mean per unit change in the predictor variable $\boldsymbol{x_i}$. In other words, increasing $\boldsymbol{x_i}$ by one unit is associated with an increase of $\boldsymbol{\beta_i}$ in the log of the mean. The exponentiated regression coefficient $\mathbf{e^{\beta_i}}$ represents a multiplicative effect of the $\mathbf{i^{th}}$ predictor on the mean. Increasing $\boldsymbol{x_i}$ by one unit multiplies the mean by a factor $\mathbf{e^{\beta_i}}$ **(**Bryan and Manfred, 2008**)**.

3.5.1. Poisson Random Effects Model

In spatial models we have a count of people y_i in area i who have a particular disease (TB in this case) in a given period of time. We can model y_i as a Poisson random variable with mean μ_i. Each area has its own value μ_i, implying that the mean can be considered as a spatially varying process as formulated in Equation 9 above. The model is established as a linear predictor of $\log\mu_i$. This predictor is usually a linear function of the logarithm of the explanatory variables in case of TB case loading model. There is some random variation in the data which can't be explained by the model above. So we allow for some normal random effects ($\varepsilon_i's$) in the linear predictor such that:

$$Log\mu_i = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon_i \qquad [12]$$

where β_i is the Poisson regression coefficient for the i^{th} explanatory variable x_i. The coefficients are estimated based on the value of observed data. Generally, we can write this formula in matrix form as:

$$Log\mu_i = x_i'\beta + \varepsilon \qquad [13]$$

The vector $\boldsymbol{\varepsilon} = \{\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \dots, \boldsymbol{\varepsilon_n}\}^{\mathbf{T}}$ of spatial effects could be assumed to follow a multivariate normal distribution and in particular, $\varepsilon \sim N(0, \delta^2 W^{-1})$ where W is the proximity or the weight matrix. It is often convenient to interpret ε_i representing a residual between area variability due to unknown or unmeasured risk factors. In such cases spatial effects are represented by using a Gaussian Conditional Autoregressive Model (Prince, 2010)

3.5.2. An overdispersed Poisson regression model

In the standard Poisson regression model, the mean and variance of y_i are assumed to be equal. This assumption is however, violated in many applications, as events often occur in clusters. The standard Poisson regression would have been misspecified under the mean variance equality assumption (Sturman, 1999, Cameron and Trivedi, 1986). It is possible to account for overdispersion with respect to the Poisson model by introducing a dispersion parameter $\emptyset$ into the relationship between the variance and the mean and use alternative models such as over dispersed Poisson and Negative Binomial Regression models (Winkelmann, 2003).

In an over dispersed model we assume that the variance is a linear function of the mean (McCullagh and Nelder, 1989):

$$Var(y_i) = \phi(\mu_i) \qquad [14]$$

where ϕ denotes the dispersion parameter. Equation 14 is known as the linear variance function since in it the variance of y_i increases as a linear function of μ_i. McCullagh and Nelder

(1989) suggested estimating the dispersion parameter ϕ as the ratio of Pearson's chi square to its associated degree of freedom. In other words, the overdispersion parameter ϕ of Equation 14 can be estimated by:

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \qquad [15]$$

where $\hat{\mu}_i = e^{x_i'\hat{\beta}}$ and $\hat{\beta}$ is the maximum likelihood estimator of β under the null hypothesis of overdispersed Poisson model. We call this Poisson regression model with correction to the estimated variance of the coefficients as overdispersed Poisson regression model. The assumptions of overdispersed Poisson regression model are the following**:**

- ✓ Logarithm of the disease rate changes linearly with equal increments in the explanatory variables.
- ✓ Changes in the rate from combined effects of different explanatory variables or endogenous variables are multiplicative.
- ✓ At each level of the covariates, the value of the dependent variable has larger variance than mean.
- ✓ An assumption of Poisson regression which says observations are independent is violated.
- ✓ $E(\varepsilon) = 0$
- ✓ $cav(\varepsilon) = \sigma^2 W^{-1}, \sigma^2 > 0,$ where W^{-1} is the inverse of the weight matrix.

3.5.2.1. Maximum Likelihood Estimation

Let yi be an observed count and assumed to be drawn from the Poisson distribution with a conditional mean μ_i on a given explanatory variable x_i , for say area i. Then, the density function of y_i can be expressed as (McCullagh and Nelder, 1989)**:**

$$f(y_i \mid x_i) = \frac{e^{-\mu_1} \mu_i^{\,y_i}}{y_i!} \qquad [16]$$

The model parameters $\mu_i = e^{x_i'\beta}$, $i = 1,2,3,\ldots,k$ (k explanatory variables) are estimated by the maximum likelihood method. The log likelihood is given by the equation:

$$\ln L = \sum_{i=1}^{n} \left[-\mu_i + y_i x_i' \beta - \ln y_i! \right] \qquad [17]$$

$$\hat{\beta} = \sum_{i=1}^{n} (-\mu_i + y_i) x_i \qquad [18]$$

For standard Poisson regression model, the variance covariance matrix is computed as the inverse of the Hessian given by:

$$H = \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} x_i x'_i \, \mu_i \qquad [19]$$

For overdispersed Poisson regression model, the variance covariance matrix of $\hat{\beta}$ is obtained by multiplying Equation 19 by estimated dispersion parameter and given by (McCullagh and Nelder, 1989).

$$Var\left(\hat{\beta}\right) = \left(\left[-\sum_{i=1}^{n} \left[x_i x'_i \, e^{x_i' \beta} \right] \right]^{-1} \hat{\phi} \right) \qquad [20]$$

Once we obtain the parameter estimates, i.e. estimates of β, we can calculate the conditional mean $\hat{\mu}_i = \mathbf{e}^{\mathbf{x_i'}\hat{\beta}}$, which gives us the expected relative risk per areal unit.

The 95% confidence interval of **β** is given by

$$\left(\hat{\beta} \pm z_{\alpha/2} . Se\left(\hat{\beta}\right) \right) \qquad [21]$$

where is $Z_{\alpha/2}$ a critical value on the standard normal distribution. For a given predictor variable with a level of 95% confidence, we would say that we are 95% confident that upon repeated trials, the confidence interval would include the true population Poisson regression coefficient.

3.5.2.2. Method of goodness of fit for overdispersed Poisson regression model

Having fitted a statistical model to the data, diagnostic tests are needed to assess the fit of the model. Goodness of fit tests use the properties of a hypothesized distribution to assess whether or not observed data are generated from a given distribution (Read and Cressie, 1988). As we had mentioned, a major assumption in an overdispersed Poisson regression model is the variance of the count data are larger than the mean. Here the diagnostic tests are concerned with checking for this assumption. The most well known goodness of fit test of statistics used are the Deviance function and Pearson's chi square.

Pearson's Chi square Statistic

One assumption of an overdispersed Poisson regression model is the variances are larger than the mean. In practical terms we assume proportionality, that is $Var\left(y_i\right) = \phi\mu_i$. The statistical hypothesis associated to this is:

$\mathbf{H_0}$: $\phi = 1$

$\mathbf{H_1}$: $\phi > 1$

The appropriate test of statistic for this hypothesis is Pearson's chi square

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\sigma}_i} \ , \qquad [22]$$

where y_i the observed data, μ_i is the true mean from the model, and δ_i is the error and is usually represented by the standard deviation of y_i. In large samples the distribution of Pearson's statistic is approximately chi squared with $n - p$ degrees of freedom where n the number of observations and p is the number of parameters. Several authors have proposed estimating ϕ using Pearson's chi squared statistic divided by its degree of freedom.

$$\hat{\phi} = \frac{\chi^2}{n - p} \qquad [23]$$

Nelder (1989) interpreted the dispersion parameter in the analysis of parameter estimates as, when $\hat{\phi} = 1$, we have a standard Poisson regression model and when $\hat{\phi} > 1$, we have an overdispersed Poisson regression model. In such case the data fits the model well (TB clustering). However, in case of model fitting for case clustering the ratio is quite larger than one. If there is no overdispersion that represents standard Poisson regression model, the Pearson's chi square will roughly equal with difference between the numbers of observations in the data set minus the number of parameters in the model.

Deviance function for testing over dispersed Poisson regression model

The goodness of fit of Poisson regression model can also be evaluated using the deviance function (McCullagh and Nelder, 1989). A measure of discrepancies between observed and fitted values is the deviance. The deviance, denoted by G^2, is calculated as twice the difference

between the log likelihood under the maximum model and the log likelihood under the reduced (or unsaturated) model and is given by (Wood, 2002):

$$G^2 = 2\sum_{i=1}^{n}[y_i \ln(\frac{y_i}{\mu_i}) - (y_i - \mu_i)] \quad [24]$$

The hypothesis is:

$H_o : \phi = 1$ (The fitted model is a standard Poisson regression model)

$H_1 : \phi > 1$ (The fitted model is an overdispersed Poisson regression model)

For large samples the distribution of the deviance is approximately a chi squared with $n - p$ degrees of freedom. The goodness of fit tells us that if the ratio of the deviance to its degree of freedom is larger than 1 an overdispersed Poisson regression model is adequate or the data fits the model well.

3.6. Negative Binomial Regression Model

Negative binomial regression model, is another alternative method used to deal with overdispersion in count data such as TB case loading frequencies (Bauer et al, 2007).

The condition that the mean and variance must be equal is nullified in the presence of overdispersion in count data. Count data often exhibit overdispersion with variance larger than the mean. Another count model which allows for overdispersion is the negative binomial model. The negative binomial distribution can be derived from the Poisson when the mean parameter is not identical for all members of the population, but itself is gamma distributed. Unlike the Poisson distribution, the negative binomial adds a quadratic term to the variance in representing an over dispersion. (Prince, 2010).

The probability density function of negative binomial random variable is given by**:**

$$f(y_i;\mu_i;\phi)=\frac{\Gamma(\phi+y_i)}{\Gamma(\phi)\Gamma(y_i+1)}\left(\frac{\phi}{\mu_i+\phi}\right)^{\varphi}\left(\frac{\mu_i}{\mu_i+\phi}\right)^{y_i} \quad [25]$$

The mean and variance of negative binomial regression model are given by

$$E(y_i;\mu_i;\phi)=\mu_i$$

$$Var(y_i;\mu_i;\phi)=\mu_i+\frac{\mu_i^{\,2}}{\phi}$$

where, $\mu_i=e^{x'_i\beta}$, $\Gamma(.)$ is a gamma function and ϕ is the dispersion parameter that must be estimated.

Fitting negative binomial regression model is very similar with overdispersed Poisson regression model. That is, the log of the mean μ_i is a linear function of explanatory variables.

$$\log\mu_i=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_nx_k+\varepsilon_i \quad [26]$$

This implies that μ_i is an exponential function of independent variables and the denotation of the coefficients, explanatory variables and assumptions are the same as in over dispersed Poisson regression model. The relationship between the variance and the mean of negative binomial distribution is presented by

$$Var(y_i)=\mu_i+\phi^{-1}\mu_i^{\,2} \quad [27]$$

The dispersion parameter is usually assumed to be fixed and can be estimated from observed data using the method of moments by using the maximum likelihood. For an overdispersed count variable, negative binomial regression model is better than Poisson Regression. If $\phi^{-1}\to 0$ in value (when ϕ gets large) we obtain the fitted model is a standard Poisson regression model. If $\phi^{-1}>0$, then the variance is larger than the mean. Thus, the negative binomial distribution is overdispersed relative to the Poisson. One important characteristic of the negative binomial

distribution is that it naturally accounts for overdispersion due to its variance always being greater ($\emptyset^{-1} > 0$) than the variance of a Poisson distribution with the same mean μ_i (Zhang et al, 2007).

3.6.1. Maximum Likelihood Estimation Method

As in Poisson regression model, the negative binomial regression coefficients $\beta_0.\beta_1, \beta_2, \ldots$ are estimated by maximum likelihood method. The estimation of the model parameters can be done by minimizing the log likelihood function. For the negative binomial distribution, the log likelihood is given by the equation (Bauer et al 2007).

$$\ln L(\phi;\beta) = \sum_{1=1}^{n}\left\{\sum_{j=1}^{y-1}\ln(j+\phi) - \ln y_i! - (y_i+\phi)\ln(1+\phi^{-1}\mu_i) + y_i \ln \phi^{-1} + y_i \ln \mu_i\right\} \quad [28]$$

where $\mu_i = e^{x'_i\beta}$, from this log likelihood function it is possible to obtain the parameter estimates.

3.6.2. Test for Goodness of Fit

In the presence of overdispersion an alternative approach to model is negative binomial regression model. For negative binomial distribution the major assumption considered is

$$Var(y_i) = \mu_i + \phi^{-1}\mu_i^{\,2},\ E(y_i) = \mu_i \quad [29]$$

where $\emptyset^{-1} \geq 0$ indicates dispersion parameter. When $\emptyset^{-1} \to 0$, the negative binomial distribution reduces to Poisson. Thus, in testing an overdispersion the null hypothesis is defined as

$$\mathbf{H_0}: \ \phi^{-1} \rightarrow \mathbf{0}$$

and the alternative hypothesis is

$$\mathbf{H_1}: \phi^{-1} > \mathbf{0}.$$

The null hypothesis defined indicates that the fitted model is a standard Poisson regression model. The appropriate test statistics are the deviance function and the Pearson's chi square as discussed in an overdispersed Poisson regression model. A decision about whether the Poisson form is appropriate can be based on one of several statistics. Thus, Pearson's χ^2 statistic is given as

$$\chi^2(\mu_i : n) = \sum_{i=1}^{n}\sum_{i=1}^{n} \frac{[y_i - \mu_i]^2}{\mu_i + \frac{\mu_i^2}{\phi}} \qquad [30]$$

A value of the Pearson's chi square greatly in excess of $\chi^2_{(n-p)}$ suggests that the model is overdispersed due to non Poisson form. Thus, when Pearson's chi squared statistic is divided by degree of freedom

$$\frac{\chi^2(\mu; n)}{n - p} \qquad [31]$$

is larger than 1, this indicates overdispersion. Likewise, the deviance model based on the definition of the scaled deviance (Wood, 2002), the G^2 statistic for a negative binomial model is given by

$$G^2(\mu_i : \phi : n) = 2\sum_{i=1}^{n}\left[\phi \log\left(\frac{\mu_i + \phi}{y_i + \phi}\right) + y_i \log\left(\frac{y_i(\mu_i + \phi)}{\mu_i(y_i + \phi)}\right)\right] \qquad [32]$$

A value of the deviance greatly in excess of $\chi^2_{(n-p)}$ suggests that the model is an overdispersed due to non Poisson form. Thus, when deviance is divided by degrees of freedom

$$\frac{G^2(\mu_i;\phi;n)}{n-p} \qquad [33]$$

is larger than one, this is an indication of overdispersion. Pearson's chi- square and deviance statistic should be approximately chi-square distributed with $n-p$ degree of freedom.

Test of overdispersion

Deviance and Pearson's Chi-square divided by degree of freedom are used to detect overdispersion in Poisson regression. Values greater than 1 indicate overdispersion that is the true variance bigger than the mean. We can test the significance of overdispersion with a Likelihood Ratio Test which follows chi square ($\chi^2_{(1-2\alpha,1d.f)}$) distribution with 1 degree of freedom based on Poisson and Negative Binomial distributions. This test tests equality of the mean and the variance imposed by the Poisson distribution against the alternative that the varianc e exceeds the mean. For the negative binomial distribution, $Var(y_i) = \mu_i + \phi^{-1}\mu^2{}_i$ the negative binomial distribution reduces to the Poisson when $\phi^{-1} \to 0$.

Therefore, in testing overdispersion the hypothesis is given by:

$$H_0 : \phi^{-1} \to 0$$
$$H_1 : \phi^{-1} > 0$$

The Likelihood Ratio Test statistic for this hypothesis is given as:

$$LR = -2[LL(Poisson) - LL(negativebinomial)]$$

Reject H_0 **:if** $LR > \chi^2_{(1-2\alpha,1df)}$

In all the above tests whenever overdispersion exists in the Poisson regression model (i.e. H_0 : is rejected), it is recommended to use negative binomial regression model.

3.7. Methods of model selections

Statistical comparisons between Poisson and negative binomial regression models confirm that in most cases the negative binomial regression model better represents observed counts that are overdispersed than Poisson regression model (Hausman et al. (1984)).

Negative binomial regression is the extension of Poisson with modification in variance assumption and will be equal to Poisson regression when the dispersion parameter gets large or $\emptyset^{-1}$ is equal to zero. This important fact provides a possibility to make comparison between Poisson and negative binomial regression models. First of all, we can look at the value of chi square statistic of dispersion parameter to assess the significance of over dispersion. Then, a likelihood ratio (LR) test, which follows chi square distribution with 1 degree of freedom, between two regressions can be used to determine the preferred model for the data. The likelihood ratio test is calculated as minus twice the difference between the log likelihood under the maximum model (Poisson) and the log likelihood under the reduced (or unsaturated) model (negative binomial):-

$$LR = -2(\log Poisson - \log Negativebin) \qquad [34]$$

A large value of log likelihood indicates that the model is a preferable model. Generally, in choosing between two models the information-based criteria mainly AIC and BIC are also used.

3.7.1. Akaike Information Criterion (AIC)

The AIC is another measure of fit that can be used to assess models. This measure also uses the log likelihood, but adds a penalizing term associated with a number of variables. It is well known

that by adding variables, one can improve the fit of models. Thus, the AIC tries to balance the goodness of fit versus the inclusion of variables in the model. The AIC is given by

$$AIC = -2\ln L + 2p \qquad [35]$$

where p is the number of unknown parameters included in the model and $\ln L$ is the log likelihood described in Equation 28. Smaller values indicate the best models to be selected.

3.7.2. Bayesian Information Criterion (BIC)

Similar to AIC, BIC also employs a penalty term associated with the number of parameters p and the sample size n. This measure is also known as the Schwarz Information Criterion and is computed as

$$BIC = -2\ln L + p\ln n \qquad [36]$$

Again, a smaller value of BIC suggests a better fit.

CHAPTER FOUR

RESULTS AND DISCUSSION

In this study, both overdispersed Poisson and negative binomial regression models are used to see the significance and type of relationship that exists between the dependent and independent variables after the nature of the distribution of TB cases are determined by using Global Moran's I and Geary's C statistics. In addition to that, Moran scatter plots are also used to identify local spatial clustering mainly to identify clustering of high values and low values.

The explanatory variables included in this study are expected to have significant effect on the dependent variable TB case loading. The explanatory variables considered in this study are population density, number of health centers and prevalence of HIV cases.

4.1. Global indicators of spatial autocorrelation

Moran's and Geary's indices of spatial autocorrelation were used to measure the degree of correlation of TB disease among neighboring woredas in North Shoa Zone based on empirical relation of contiguity weight matrix (Queen's method).

Under the null hypothesis that there exists no spatial autocorrelation (randomness) between neighboring woredas Moran's *I* and Geary's *C* coefficients were computed to test the clustering of TB cases in the entire study region at $\alpha = 0.05$ level of significance. The elements of the contiguity weight matrix were calculated by using Queen's method (Figure 1 and Table 1, Annex 1) because Queen's method considers common edge and/ or vertex in defining the spatial dependence.

Results of Moran's index in the Table 1 below show that there is a significant positive spatial autocorrelation ($I = 0.635$) between neighboring woredas of TB cases in North Shoa Zone. This means that, the spatial distribution of TB cases is clustered (globally) in the study area.

Table 1. Results of global Moran's I and Geary's C

Measures	Observed	p-value
Moran's *I*	0.635	0.0001
Geary's *C*	0.485	0.095

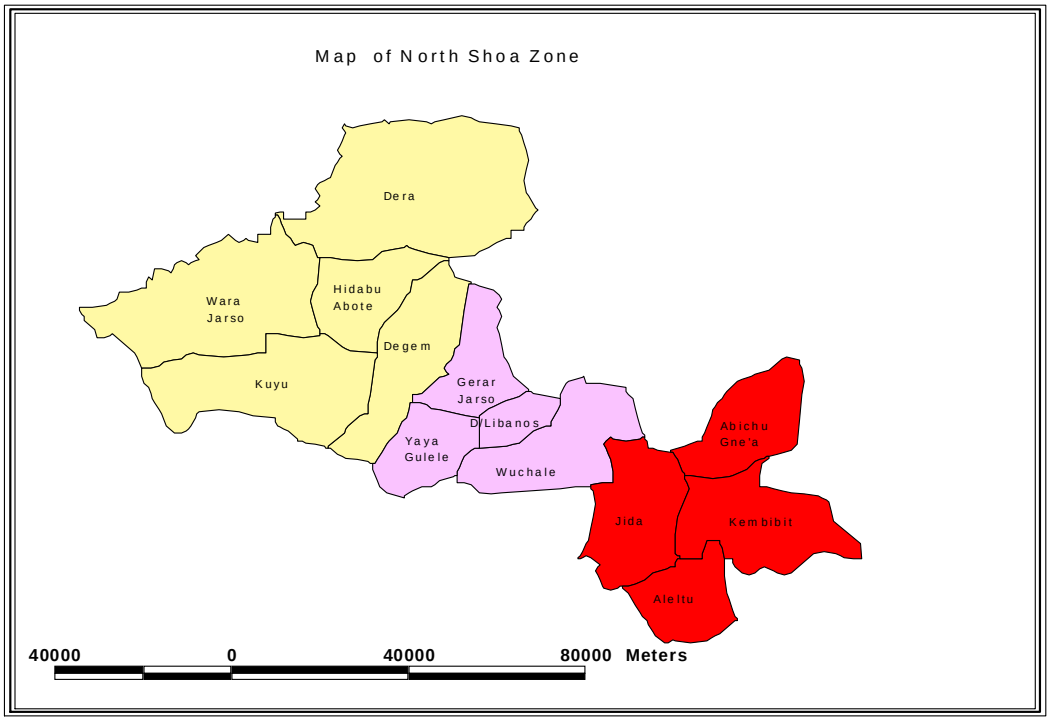


Figure 1.Map of North Shoa Zone

The results also indicate the presence of strong local patterns in the spatial distribution of TB that needs to be further explored by using local indicators of spatial autocorrelation (Figure 2).

4.2. Local indicators of spatial autocorrelation

Moran scatter plot is one way of identifying local spatial clustering of TB in woredas which could be identified as hot spots and cold spots of TB. The map of North Shoa Zone drawn through ARCGIS software is used as an input to investigate the composition of the global Moran statistic. A local Moran statistic can be calculated for each polygon (this is known as a local indicator of spatial association, or LISA statistic). From Figure 2 below one can easily realize the spatial dependence of woredas that are nearest to one another.

Figure 2 depicts the result of Moran scatter plot of TB diseases in North Shoa Zone; the local Moran statistic for each variable is plotted in the form of a Moran scatter plot. In this Moran scatter plot, the horizontal axis specifies the observed values of TB cases while the vertical axis specifies the weighted average of neighboring values of TB cases. The first and third quadrants indicate the presence of TB clustering of similar values while the others are used in representing TB clustering of dissimilar values**.**

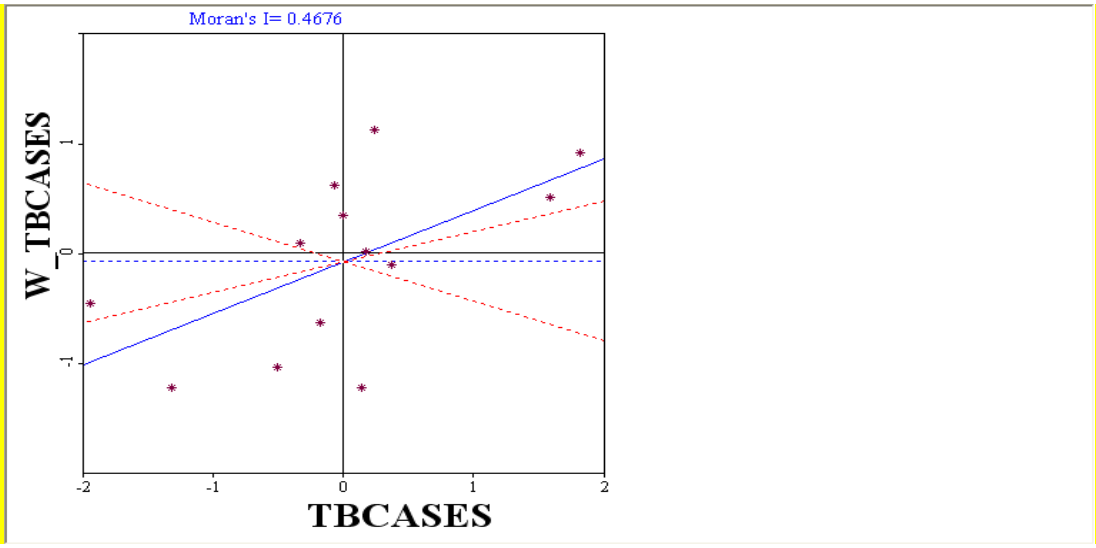


Figure 2. Moran scatter plot

Clearly, from the first quadrant (upper right) of Moran scatter plot we can understand that in five woredas the distribution of TB diseases are highly clustered. This result indicates that in these

five woredas, there is high TB clustering of similar values (hot spots). From the 3rd quadrant (lower left) we see that the distribution of TB diseases in four woredas is less clustered. This indicates that in the four woredas the distribution of TB diseases are cold spots (low low).

On the other hand, as it is seen from the second and fourth quadrant (lower right and upper left) of the Moran scatter plot that in four woredas there is TB clustering of dissimilar values either high low or low high.

However, identification of woredas for the presence of significant TB clustering is done based on cluster map and depicted in Figure 3 as hot spots, cold spots and clustering of dissimilar values.

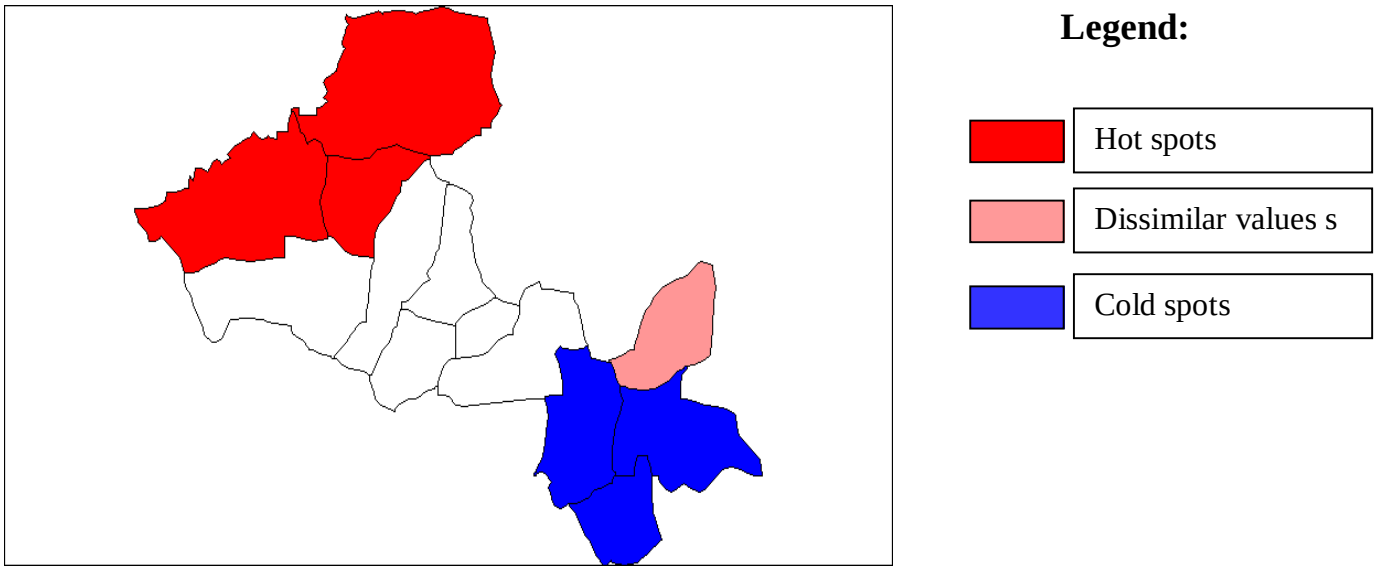


Figure 3. Significant TB clustering

In Figure 3, the red color indicates the presence of significant hot spots (high TB clustering) in five woredas as indicated in Moran scatter plot while the blue color indicates the presence of significant cold spots (low TB clustering) in the four woredas. The map reveals that in Dera, Wera Jarso and Hindhibu Abote woredas there is a significant TB clustering of high values (hot spots). In Jida, Aleltu and Kembibit a significant TB clustering of low values (cold spots) is observed. On the other hand, only in Abicho Nya'a the presence of significant TB clustering is observed in all other woredas found to be dissimilarly clustered.

Table 2. TB case loading in North Shoa Zone, Oromia Region, 2008.

Name of Woreda	Population Size (year 2007)	Population density	All forms of TB cases
Abicho Nya'a	73,269	183.63	368
Aleltu	53,361	156.48	90
Dagem	98,208	476.74	260
Debre Libanos	46,850	92.59	182
Dera	181,661	535.79	370
Girar Jarso	94,785	239.40	270
Hindhibu Abote	85,691	238.69	256
Jida	53,769	72.46	142
Kembibit	74,247	252.54	239
Kuyu	126,546	248.13	344
Wera Jarso	148,771	287.76	274
Wuchale	97,470	219.52	249
Yaya Gulale	54,958	45.87	110
Grand	**1,263,455**		**3,154**

We would like to indicate that the population size for Girar Jarso includes the population of Fiche town. Also, based on the result of Table 2, the population density in Dera is larger as compared to all other woredas. On the other hand, Yaya Gulalle has least population density.

Generally speaking, since the distribution of TB in North Shoa Zone is clustered, the usual assumption of randomness and mean variance equality does not hold. Therefore, it is not appropriate to fit a standard Poisson regression model.

In such cases overdispersed Poisson regression and negative binomial regression models are applied. Since there is statistically significant overdispersion in the data of this study, the analysis is done by applying both an overdispersed Poisson regression model and negative binomial regression model.

4.3. Poisson Regression Analysis

As mentioned in Section 4.2 above, both an overdispersed Poisson and negative binomial regression models were used for analyzing TB case loading. Eventhough both models are recommended when cases are overdispersed, they are different in the assumption about mean variance equality. In an overdispersed Poisson regression model the basic assumption is that the variance is a linear function of the mean while in a negative binomial regression model it is a quadratic form of the mean.

4.3.1. Overdispersed Poisson Regression Estimator

In this section we identify explanatory variables that have significant effect on TB case loading observed in North Shoa Zone.

The analysis for an overdispersed Poisson regression model was started by testing the significance and association of each explanatory variable could have with the dependent variable, namely with TB case loading.

For this purpose, a chi square test is used in both overdispersed Poisson regression and negative binomial regression models. In the data fitting process all explanatory variables were entered into regression function given by**:**

$$Log\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where μ indicates relative risk TB cases in each woreda β_0 is the intercept, $\beta_1, \beta_2, \beta_3$ are regression coefficients for the respective explanatory variables x_1, x_2, x_3. The variables

x_1, x_2, x_3 respectively are population density, number of health centers and prevalence of HIV cases.

The relation between TB cases loading with their corresponding explanatory variables (covariates) was analyzed using both regression models. For each covariate, we used an overdispersed Poisson regression and negative binomial regression models analysis that contains a single dependent variable in order to have an idea about each explanatory variable. We presume that the relationship between the response variable and the explanatory variables (predictors) are either negative or positive based on their estimated regression coefficients as it can be seen from Table 3 below.

Table 3. Results of the overdispersed Poisson regression model

Parameter	Estimate	Wald Chi square	Pr>Chisq
Intercept	4.4862	57.88	<0.0001
Population density	0.3200	54.94	<0.0001
Health center	-0.0134	5.55	0.0185
HIV	0.0150	17.35	0.0007

**** At $\alpha = 0.05$ level of significance**

The Chi square test results indicate that all explanatory variables are found to be statistically significant in determining TB distribution observed in North Shoa Zone. However, as clearly presented in Table 3 above TB case loading due to population density is found to be higher than other explanatory variables.

In general, the relative risk of TB cases is found to be high in areas with population density and prevalence of HIV cases. However, since all explanatory variables are found to be significantly associated with the dependent variable, as per the chi square test result, they are considered in an overdispersed Poisson regression analysis. Accordingly, the fitted model is given by:

$$Log\mu = 4.4862 + 0.3200x_1 - 0..0134x_2 + 0.0150x_3$$

The results show that the number of health centers per woreda (negative coefficient) have an effect to decrease TB case loading whereas all explanatory variables have effects to increase (positive coefficients) TB cases.

4.3.2. Results of goodness of fit for an overdispersed Poisson regression model

Tests for the adequacy of the model and the significance of the predictor variables can be established by using deviances and Pearson's chi square statistics. The results are presented in Table 4 below. From the results it can be concluded that the model fits the data well (measures for both deviances and Pearson's chi square statistics are 73.8468 and 72.8243 respectively). Furthermore, the values of deviances and Pearson's chi square divided by the degrees of freedom are larger than one supporting the presence of overdispersion in TB clustering.

Table 4. Results of goodness of fit test

Criterion	Degrees of freedom (d.f)	Value of statistic	Value of dispersion parameter = (Values of statistic ⎪ d.f)
Deviance	9	73.8468	8.2052
Scaled deviance	9	73.8468	8.2052
Pearson chi-square	9	72.8243	8.0916
Scaled Pearson chi-square	9	72.8243	8.0916
Log likelihood		14664.8141	
Full log likelihood		-84.1996	
AIC(smaller is better)		176.3991	
AICC(smaller is better)		181.3991	
BIC(smaller is better)		178.6589	

4.4. Negative Binomial Regression Estimator

An alternative method for analyzing the relationship between predictor variables and response variable is the negative binomial regression. For each covariate, one can determine the significant effect on the response variable. The results shown in Table 5 indicate the presence of either positive or negative relationship between dependent and explanatory variables. The values of the coefficients suggest positive relationship between TB case loading and independent variables except health centers which have a negative impact.

The result obtained from this model also indicates that population density, number of health centers and prevalence of HIV significantly affect TB case loading.

Table 5. Results of the negative binomial regression model.

Parameter	Estimate	Wald Chi Square	Pr>ChiSq
Intercept	4.3768	54.36	<0.0001
Population density	0.3600	34.98	<0.0001
Number of health center	-0.0722	11.98	0.0016
Prevalence of HIV cases	0.0190	4.04	0.0443

**** at $\alpha = 0.05$ level of significance**

The values of dispersion parameter shown in Table 6 below indicate that the model fitted by negative binomial regression model is overdispersed. In addition, the values of dispersion parameter are all greater than zero indicating that the mean of TB cases is less than the variance.

For further analysis, using χ^2 test shows that all explanatory variables are found to be significantly associated with the dependent variable (TB case loading).

Accordingly, the fitted model is given by:

$$Log\mu = 4.3768 + 0.3600x_1 - 0.0722x_2 + 0.0190x_3$$

Where μ indicates TB case loading per woreda. The constant term (intercept) is estimated to be 4.3768, x_1 stands for the population density, x_2 stands for the number of health centers, x_3 stands for the prevalence of HIV cases.

4.4.1. Results of goodness of fit for the negative binomial regression model

We can test the adequacy of this model and the significance of the predictor variables using deviances and Pearson's chi square test statistics just as in overdispersed Poisson regression model discussed in the previous section. The results of the tests are presented in Table 6 below, and these indicate that the fitted model is good.

Table 6. Results of goodness of fit test.

Criterion	Degrees of freedom (d.f)	Value of statistics	Value of dispersion parameter = (Values of statistic ⎪ d.f)
Deviance	9	13.3159	1.4795
Scaled Deviance	9	13.3159	1.4795
Pearson Chi-Square	9	13.2797	1.4755
Scaled Pearson chi-Square	9	13.2797	1.4755
Log-Likelihood		14683.1464	
Full Log Likelihood		-65.8673	
AIC(smaller is better)		141.7345	
AICC(smaller is better)		150.3059	
BIC(smaller is better)		144.5593	

Specifically, the values of the deviance and Pearson's chi square are nearly the same (13.3159 and 13.2797 respectively). Furthermore, the ratio of values of deviance's and Pearson's chi square to their corresponding degree of freedom are larger than zero, indicating the presence of overdispersion in TB clustering.

4.4.2. Results of test of overdispersion

For the Poisson model, the Pearson Chi-square and deviance values divided by the degrees of freedom are sufficiently larger than 1. This is a possible indication that the fit is overdispersed; however, this is also justified by applying a formal statistical test of dispersion.

Criterion	**Estimate**	**Poisson model**	**Negative binomial model**
Deviance	Value/d.f	8.2052	1.4795
Pearson's Chi-square	Value/d.f	8.0916	1.4755
Log likelihood	Value	14664.8141	14683.1464

To carry out the **LR test** for significance of overdispersion, that is to test the hypothesis:

$$H_0 : \phi = 1$$

$$H_1 : \phi > 1$$

The result obtained from **-2(LL(Poisson) – LL(negative binomial)** is **36.6646**, which corresponds to P-value = **0.0001**. Hence, we reject H_0 and conclude that the mean and variance are not equal as a result the assumption of standard Poisson regression model has to be abandoned. This clearly, shows that presence of significant overdispersion.

4.5. Comparisons of an overdispersed Poisson and negative binomial regression models

In Table 7, three different methods are used to compare overdispersed Poisson and negative binomial regression models: log likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC). The results obtained indicate there is observed difference in values between the two models. Since negative binomial regression model has smaller value in AIC and BIC. Consequently, we conclude that in this study negative binomial regression model is better than Poisson for modeling an overdispersed data.

Table 7. Results of comparison of Poisson and negative binomial regression models

Methods of comparisons	Overdispersed Poisson regression model	Negative binomial regression Model
Log Likelihood Ratio test	14664.8141	14683.1464
Akaike Information Criterion	176.3991	141.7345
Bayesian Information Criterion	178.6589	144.5593

CHAPTER FIVE
SUMMARY AND CONCLUSION

Geographical clusters of TB cases were identified through exploratory spatial data analysis, using Global Moran's *I*, Geary's *C* and also local indicators. The results obtained reveal that the distribution of TB in North Shoa Zone is clustered indicating overdispersion. Furthermore, the Moran scatter plot also depicts the presence of strong local spatial clustering of high values in five woredas (hot spots) of which in Dera, Hindhibu Abote and Wera Jarso woredas the most significant TB clustering of high values (hot spots) was observed. Significant TB clustering of low values were observed in Aleltu, Jida and Kembibit woredas from among the four woredas shown in Moran scatter plot. On the other hand, only in Abichu Nya'a there is significant TB clustering of dissimilar values.

Based on the Poisson and negative binomial regression population density, number of health centers, and prevalence of HIV cases were identified to be significantly associated with TB case loading.

For count data with the evidence of overdispersion, negative binomial regression model is preferred to the Poisson regression model.

5.1. RECOMMENDATIONS

This paper has endeavored to analyze the spatial distribution of TB in the North Shoa Zone. The results of this study demonstrate that in the study area TB case loading pattern varies from woreda to woreda and is clustered. The presence of spatial dependence between woredas was also established. Based on the results obtained from the fitted models the following recommendations can be made.

- Pay special attention in combating HIV/AIDS epidemy.
- Expand and strengthen medical facilities especially in woredas identified as hot spots.
- Provide intensive family planning advise especially in woredas identified as hot spots.

6. REFERENCES

1. Abera, B., Kate, F., Zelalem, H. and Andrew, F. (2009)**:** The association of TB with HIV infection in Oromia Regional National State, Ethiopia. *Ethiopian Journal of Health Development 23(1), 63-66.*
2. Alpharetta, A. (2008): Count Data Models in SAS. Statistics and Data Analysis. *Journal of Applied Econometrics 9, 1-12.*
3. Anselin, L. (1995)**:** Local Indicators of Spatial Autocorrelation (LISA). *Geographical Analysis 27(2), 93-115.*
4. Anselin, L. (1996): The Moran Scatter plot as an ESDA (explanatory spatial data analysis) Tool to assess local instability in spatial association. *Spatial Analytical Perspectives on GIS 28,111-125.*
5. Anselin, L. and Hudak, J. (1998): Spatial econometrics in practice and reviews of soft ware options. *Regional Science and Urban Economics 22, 509-536.*
6. Asnakew, K., Yeshiwondim, M., Sucharita, G., Afework, T. and Dereje, O. (2007): Spatial analysis of malaria incidence at the village level in areas with unstable transmission in Ethiopia. *International Journal of Health Geographics 8, 1-11.*
7. Awash, T., Ingrid, P. and Borrel, N. (2009)**:** A Temporal-Spatial Analysis of Malaria Transmission in Adama (Nazereth), Ethiopia. *American Journal of Tropical Medicine and Hyg.81 (6), 944-949.*
8. Barrette, F. and Finke, K. (1792)**:** Map of human disease and the first world disease map. *Social Science Medicine 80, 915-921.*
9. Bauer .L., Greibe, P., Hua, L. and Liang, L. (2007): Statistical Models of Accidents on interchange ramps and speed change lines, FHWA- RD – 97-106, U.S. Department of Transportation.
10. Beyer, N. and Tatley, M. (1996)**:** The use of a geographical information system (GIS) to evaluate the distribution of TB in a high incidence community. *South Africa Medical Journal 86, 41-44.*
11. Bryan, T. and Manfred, M. (2008): Regression Models for Count Data. *Journal of Pediatric Psychology 33(10), 1076-1034.*

12. Cameron, A. and Trivedi, K. (1986)**:** Econometric models based on the count data, comparisons and applications of some estimators and tests. *Journal of Applied Econometrics 1, 29-53.*

13. Carl, N. (2007): Tuberculosis Incidence in Portugal and Spatiotemporal Clustering. *International Journal Health Geographics 6, 30-36.*

14. Cliff, A. and Ord, J. (1981): Spatial processes, Modeling and application. Pion, London.

15. Cressie, N. (1993): Statistics for spatial data, Wiley, New York.

16. Cressie, N. and Read, T. (1989)**:** Spatial data analysis of regional counts. *Biomedical Journal 31,699-719.*

17. Czado, C. (2008): Modeling Count Data by spatial random effects. *Statistical papers 49(3), 531-552.*

18. Diggle, P. (1993): Point process modeling in environmental epidemiology. *Statistics for the environment 23, 89-110.*

19. Federal Ministry of Health (2007): Health Promotion and TB Disease Prevention in Ethiopia.

20. Federal Ministry of Health (2008): Implementation Guidelines for TB/HIV Collaborative Activities in Ethiopia.

21. Friedrich, G. (1998)**:** Survey on cluster tests for spatial area data. *Computational Statistics and Data Analysis 31, 39-58.*

22. Gatrell, A., Bailey, T., Diggle, P. and Rowlingson, B. (1996)**:** Spatial Point Pattern Analysis and its application to medical geography. *Transactions of the institute of British Geographers, 21(1), 256-274.*

23. Goodchild, M. (1987)**:** A spatial analytical perspective on geographical information systems. *International Journal of GIS 6(5), 407-423.*

24. Hausman, J., Hall, B. and Griliches, Z. (1984)**:** Econometric models for count data with an application to the patents-R and D relationship. *Econometrica 52, 909-938.*

25. Ismael, H. (2008)**:** Prevalence study on smear positive pulmonary TB and Factors responsible for delay of patients in seeking care at health facilities Alamata Woreda, Southern Tigray. (Unpublished).

26. John, D. (1968): Environmental risk factors for Lyme disease identified with GIS. *American Journal Public Health 85,944-948.*

27. Kulldorff, M. (1997)**:** A Spatial Exploratory Statistics. Communication statistics. *Theory and Methods 26, 1481-1496.*
28. Kulldorff, M. and Nagarwalla, N. (1995)**:** Spatial Disease Clusters, Detection and Inference. *Statistics in Medicine 14, 799-810*.
29. Kunneke, M. (1999)**:** Childhood tuberculosis in an urban population in South Africa, burden and risk factors. *Archives of disease in childhood 80, 433-437.*
30. Maher, D. and Raviglione, M. (2005): Global Epidemiology of TB. *Clinics in chest medicine 26,167-182.*
31. Matthew, L. (2005)**:** The Utility of Geographical Information Systems (GIS) and Spatial temporal cluster Analysis in Tuberculosis Surveillance in Harris County, Texas. *International Journal of Health Geographics 24, 3-25.*
32. McCullagh, P. and Nelder, J. (1989)**:** Generalized Linear Models (2nd edition), London: Chapman and Hall.
33. Munche, M. (2003)**:** Tuberculosis transmission patterns in a high incidence area and spatial analysis. *International Journal of TB and Lung Disease 7,271-278*.
34. Murray, C. and Lopez, A. (1996)**:** The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries and risk factors in 1990 and projected to 2020. *World Health Organization Document. W74966L-1/1996*.
35. Nelder, D. (1989): Specification Test for Poisson Regression Models. *International Economic Review 27, 687-706.*
36. Odland, M. (1987): Spatial Autocorrelation. Sage: Beverly, CA.
37. Paez, A. and Scott, D. (2004): Spatial statistics for urban analysis and a review of techniques with examples. *Geo Journal 16, 53-67.*
38. Paulo, H. and Santos, C. (2005)**:** Spatial and temporal patterns of TB in the city of Ribeirao Prato, Brazil from 1998-2002. *Journal Brasil Pneumonia 31(1), 523-527.*
39. Pielou, E. (1961)**:** Segregation and symmetry in two species populations as studied by nearest neighbor relationships. *Journal of Ecology 49, 255-269.*
40. Porter, J. (1999)**:** Geographical information systems (GIS) and the tuberculosis DOTS strategy. *Tropical Medicine and International Health 4, 631-639.*
41. Pfeiffer, D. (2008): Issues related to handling spatial data statistics. Proceedings of the epidemiology and state veterinary programmes 23, 83-105.

42. Porto, A. (1996): Use of spatial statistics to identify and test significance in geographic disease patterns. Preventive Veterinary Medicine 11, 205-224.
43. Prince, O. (2010): Statistical methods of disease mapping. Journal of Royal Statistical Society 154(3), 421-441.
44. Pui, J., Lin, M. and Perng, H. (2006): Spatial autocorrelation analysis of health care hotspots in Taiwan. *BMC Public Health 9,464-484.*
45. Randremanana, F. (2009)**:** Spatial clustering of pulmonary tuberculosis and impact of the care factors in Antananarivo city. *Tropical Medicine and International Health 14,429-437.*
46. Read, M. and Cressie N. (1988): Goodness of fit Statistics for discrete multivariate data. Springer Verleg, New York
47. Richardson, R. (2002): Statistical Methods for geographical correlation studies. Geographical and Environmental Epidemiology, *21,181-204.*
48. Rowlingson, N. (1994)**:** Spatial point pattern analysis and its application in Geographical Epidemiology. *Trans Inst Br geographical* 21, 256-247.
49. Statistical Abstract (2007)**:** Federal Democratic Republic of Ethiopia Central Statistical Agency.
50. Snow, J. (1954)**:** GIS and spatial pattern analysis in medical geography. *Social science and medicine 50(20), 923-935.*
51. Stephan, F. (1934): Sampling errors and the interpretation of social data ordered in time and space. *Journal of the American Statistical Association 29,165-166.*
52. Sturman, M. (1999)**:** Multiple approaches to analyzing count data in studies of individual differences. The propensity for the type one errors, illustrated with the case of absenteeism prediction. *Educational and Psychological Measurement 59, 414-430.*
53. Thomas, K. (1988)**:** Spatial patterns of TB incidence. Social Science and Medicine 55(1), 7-19.
54. Tiwari, N., Adhikari, C., Tiwari, A. and Kandpal, V. (2006)**:** Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India using GIS and spatial exploratory spatial statistics. *International Journal of Health Geographic 5, 33-40.*
55. Tobler, R. (1970)**:** A computer movie simulating urban growth in the Detroit region. *Economic geography supplements 46,234-240.*

56. Touray, K., Jallow, A., Adetifa, M., Rigby, J., Jeffries, D., Cheung, Y., Donkor, S., Adegbola, A. and Hill, C. (2010)**:** Spatial analysis of tuberculosis in an Urban West African setting. *Tropical Medicine and International Health 15,664-672*.
57. WHO (2006): Global tuberculosis control program report. World Health Organization, Geneva. Switzerland.
58. WHO (2007)**:** The status of TB prevention and control activities in Ethiopia.
59. WHO (2008)**:** Global tuberculosis control program report. World Health Organization. Gambia, Africa.
60. WHO (2009)**:** Global tuberculosis control program report. World Health Organization Philippines.
61. Winkelmann, N. (2003)**:** Recent developments in count data Modeling: *Theory and application, Journal of Economic Surveys 9(1), 1-24.*
62. Wood, G. (2002)**:** Generalized Linear Accidents Models and goodness of fit testing. *Accident Analysis and Prevention 34(4), 417-427.*
*63. Zhang, V., Zhirui, N and Lord, L. (2006)***: Investigating goodness of fit test statistics for generalized linear crash models with low means.*

7. Appendix

Annex 1. Construction of weight matrix based on Queen's method

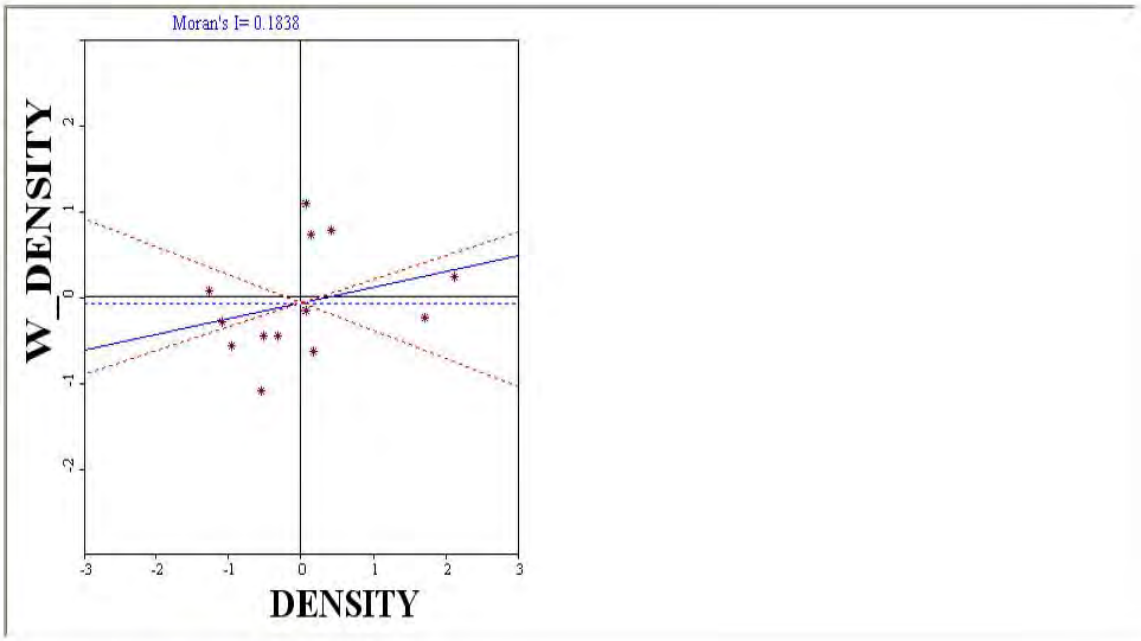
Table 1. Weight Matrix.

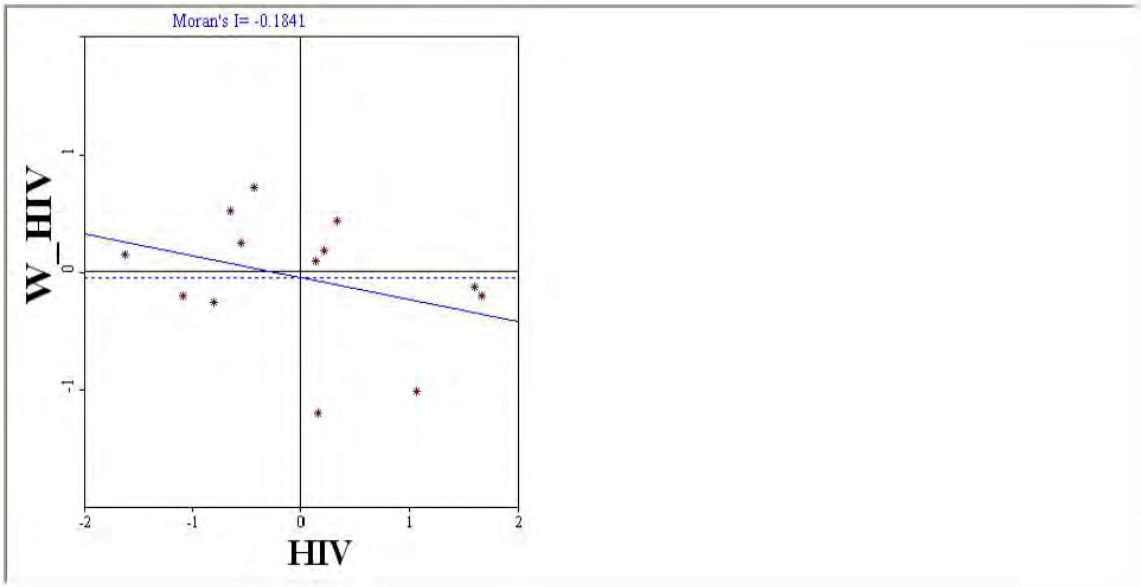
ID	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	0	0	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0	0	0	0
3	1	1	0	1	1	0	0	0	0	0	0	0	0
4	0	1	1	0	1	0	0	0	0	0	0	0	0
5	0	0	1	1	0	1	1	0	0	0	0	0	0
6	0	0	0	0	1	0	1	1	0	0	0	0	0
7	0	0	0	0	1	1	0	1	1	0	0	0	0
8	0	0	0	0	0	1	1	0	1	0	0	0	0
9	0	0	0	0	0	0	1	1	0	1	0	0	0
10	0	0	0	0	0	0	0	0	1	0	1	1	1
11	0	0	0	0	0	0	0	0	0	1	0	1	1
12	0	0	0	0	0	0	0	0	0	1	1	0	1
13	0	0	0	0	0	0	0	0	0	1	0	1	0

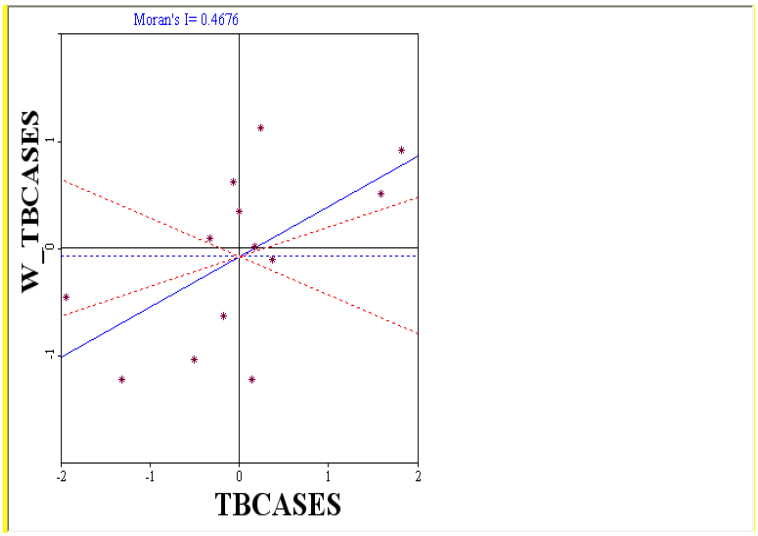
Remark: Identification numbers for woredas are given below:

Woreda	Dera	Wera Jarso	Hindhibu Abote	Kuyu	Degem	Girar Jarso	Gulale	Libanos	Wuchale	Jida	Abichu Nya'a	Kembibit	Aleltu
ID	1	2	3	4	5	6	7	8	9	10	11	12	13

Annex 2 .Graphs of Moran scatter plot for each variables






Moran's I= 0.4676
W_TBCASES
TBCASES
-2
-1
0
1
2
-1
0
1

Annex 3. SAS Output of Poisson Regression Model

The SAS System

The GENMOD Procedure

Model Information

Data Set WORK.OZONE

Distribution Poisson

Link Function Log

Dependent Variable TB case loading

Number of Observations Read 13

Number of Observations Used 13

Criteria for Assessing Goodness of Fit

Criterion	d.f	Value	Value/d.f
Deviance	9	73.8468	8.2052
Scaled Deviance	9	73.8468	8.2052
Pearson Chi-Square	9	72.8243	8.0916
Scaled Pearson X2	9	72.8243	8.0916
Log Likelihood		14664.8141	
Full Log Likelihood		-84.1996	
AIC (smaller is better)		176.3991	
AICC (smaller is better)		181.3991	
BIC (smaller is better)		178.6589	

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates (Poisson Regression Model)

Parameter	d.f	Estimate	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	4.4862	4.3267	4.6458	57.88	<.0001
Population density	1	0.3200	0.2700	0.3800	54.94	<.0001
Health center	1	-0.0134	-0.0246	-0.0023	5.55	0.0185
HIV	1	0.0150	0.0080	0.0230	17.35	0.0007
Scale	0	1.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Annex 4. SAS Output of Negative Binomial Regression model

The SAS System

The GENMOD Procedure

Model Information

Data Set WORK.OZONE

Distribution Negative Binomial

Link Function Log

Dependent Variable TB cases loading

Number of Observations Read 13

Number of Observations Used 13

Criteria for Assessing Goodness of Fit

Criterion	d.f	Value	Value/d.f
Deviance	9	13.3159	1.4795
Scaled Deviance	9	13.3159	1.4795
Pearson Chi-Square	9	13.2797	1.4755
Scaled Pearson X2	9	13.2797	1.4755
Log Likelihood		14863.1464	
Full Log Likelihood		-65.8673	
AIC (smaller is better)		141.7345	
AICC (smaller is better)		150.3059	
BIC (smaller is better)		144.5593	

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates

(Negative Binomial regression model)

Parameter	d.f	Estimate	Wald 95% Confidence Limits		Wald Chi-Square	Pr >ChiSq
Intercept	1	4.3768	4.0261	4.7275	54.36	<.0001
Population density	1	0.3600	0.2400	0.4800	34.98	<.0001
Health center	1	-0.0722	-0.1132	-0.0312	11.91	0.0016
HIV	1	0.0190	0.0100	0.0380	4.04	0.0443