60

Addis Ababa University (Since 1950)

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
COLLEGE OF NATURA SCIENCES
DEPARTMENT OF STATISTICS

SPATIAL ANALYSIS OF TUBERCULOSIS IN EASTERN HARARGE, ETHIOPIA

LETA LENCHA GEMECHU

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

June, 2013

ADDIS ABABA, ETHIOPIA

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by Leta Lencha, entitled: Spatial Analysis of Tuberculosis in Eastern Hararge, Ethiopia and submitted in Partial fulfillment of the requirements for Degree of Master of Science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner Emmanuel G/yohannes Signature _____ Date 27/06/13

Examiner Mekonnen Tadesse Signature _____ Date 27/06/13

Advisor Butte Gota Signature _____ Date 27/06/13

Chair of Department or Graduate Program Coordinator

TABLE OF CONTENTS **Pages**

ACKNOWLEDGMENTS iii
ABSTRACT iv
ACRONYMS v
CHAPTER ONE: INTRODUCTION 1
1.1. BACK GROUND OF THE STUDY 1
1.2. TUBERCULOSIS IN ETHIOPIA 3
1.3. DISEASE MAPPING AND CONCEPT OF SPATIAL DEPENDENCE 5
1.4. STATEMENT OF THE PROBLEM 6
1.5. OBJECTIVE OF THE STUDY 8
1.6. SIGNIFICANCE OF THE STUDY 8
1.7. LIMITATION OF THE STUDY 8
1.8. THESIS ORGANISATION 9
CHAPTER TWO: LITERATURE REVIEW 10
CHAPTER THREE: DATA AND METHODOLOGY 18
3.1. SOURCE OF THE DATA 18
3.3. METHODOLOGIES OF THE STUDY 20
3.3.1. SPATIAL AUTOCORRELATION 20
3.3.2. SPATIAL WEIGHTS AND NEIGHBORHOODS 22
3.3.3. GLOBAL MEASURES OF SPATIAL AUTOCORRELATION 24
3.3.4. LOCAL MEASURES OF SPATIAL AUTOCORRELATION 28
3.3.5. Testing for spatial dependence in regression residuals 31
3.4. MODELING SPATIAL DEPENDENCE 34
3.4.1. SPATIAL AUTOREGRESSIVE MODELS 35
3.4.2. MAXIMUM LIKELIHOOD ESTIMATION 38
3.5. BAYESIAN CONDITIONAL AUTOREGRESSIVE MODEL (CAR) 40
3.6. MODEL SELECTION METHODS 45
CHAPTER FOUR: RESULT AND DISCUSSIONS 48
4.1. INTRODUCTION 48
4.2. DESCRIPTIVE SPATIAL STATISTICS 48
4.3. GEO-VISUALIZATION OF THE STUDY AREA 48
4.4. SPATIAL AUTOCORRELATION 50

4.4.1. GLOBAL MEASURES OF SPATIAL AUTOCORRELATIONS 50
4.4.2. LOCAL MEASURES FOR SPATIAL AUTOCORRELATION 53
4.5. DIAGONSTIC FOR SPATIAL DEPENDENCE 56
CHAPTER FIVE: CONCLUSION AND RECOMMENDATION 70
5.1. Conclusion 70
APPENDIX 76

List of Figures pages

FIGURE 4.1 MAP OF EASTERN HARARGE ZONE 49
Figure 4.2: SNAPSHOT OF GLOBAL MORAN'S SCATTER PLOT 52
Figure 4.3 LISA CLUSTER MAP 56
Figure1: PERMUTATION TEST RESULT FOR GLOBAL MORAN'S I 76
Figure 2. BIVARIATE MORAN'S SCATTER PLOT 76

List of Tables page

Table 4.1 Global Moran's I 51
Table 4.2 Global Geary's C 51
Table 4.4: Diagnostic for spatial dependence (using row –standardized weights) 57
Table 4.5: MODEL DIAGNOSTICS SUMMARY 59
4.6. SPATIAL LAG MODEL (SLM) 59
Table 4.6: MLE results for spatial lag model 60
Table 4.7: POSTERIOR STATISTICS FOR CAR MODEL 66
Table1: Ord and Getis G* for local spatial autocorrelation 78
Table2. SOCIO-DEMOGRAPHIC VARIABLES AND TB LOAD OF WOREDAS 79
Table 3. Contiguity weigh matrix (Queen's-method) for east Hararge woredas 80

ACKNOWLEDGMENTS

First and foremost, my sincere thanks to my supervisor, **Dr. Butte Gotu,** for his constructive and excellent scientific guidance, suggestions and comments throughout my thesis work. My special thanks also go to East Hararge and Harari Regional Health bureaus, for their coordination in providing the data.

I am immensely grateful to Birhanu and Gaddisa Olan, for their efforts and supports; it's only due to your contribution that I able to use Arc GIS and shape file, which is background of my study. My heartfelt thanks go to my family for their continuous prayers and support, and in special way, I like to thank my wife, Rahel Yohannes; your motivation, support and willing comes to reality.

Once again, my thanks are endless for all individuals who support me on my thesis work in one way or other, and I give thanks to Almighty God, whose grace has sustained me to reach this level of life.

ACRONYMS

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CAR	Conditional autoregressive model
DIC	Deviance Information Criterion
EHZ	East Hararge Zone
FMOH	Federal Ministry of Health
GIS	Geographical Information system
SMR	Standardized Mortality Ratio
HIV	Human Immunodeficiency Virus
LM	Lagrange Multiplier
LRT	Likelihood Ratio Test
MCMC	Markov chain Monte Carlo
MDR	Multi drug resistant
MLE	Maximum Likelihood Estimate/Estimator
OLS	Ordinary Least Squares
RLM	Robust Lagrange Multiplier
SAR	Spatial autoregressive model
SEM	Spatial error model
SLM	Spatial lag model
TB	Tuberculosis
WHO	World Health Organization

ABSTRACT

Tuberculosis (TB) has claimed many lives throughout the history of mankind and it continues to be a global threat in the coming decades, especially in developing countries like Ethiopia. The incidence and mortality due to TB cases is not equally distributed across the globe; they vary by geographic region, subpopulation, and spread by close and prolonged contact with an infected individual.

The main aim of this study is to examine and characterize the spatial patterns of TB cases in East Hararge Zone, Oromia region, Ethiopia. The data used are obtained from Zonal Health Bureaus. Two step exploratory spatial analyses were carried out: - examining the presence of spatial autocorrelations and modeling spatial distribution of TB cases via autoregressive models. Spatial autocorrelation was investigated by global and local test statistics; *Moran's I, Geary's C, and Ord and Getis Gi** statistics. Two different spatial models; *spatial autoregressive models (SAR)* and *conditional autoregressive model (CAR)* were considered for describing spatial dependence of TB cases on local risk factors.

Test results of spatial autocorrelations revealed statistically significant clustering of fifteen woredas; of which seven woredas (located in north and eastern part) were identified as high risk areas. Spatial lag model from spatial autoregressive models and CAR model has been fitted for describing the spatial association of TB cases on explanatory variables. The result of the study suggests that TB prevalence of the study area is highly enhanced by proximity to potential affected areas, high population density, HIV prevalence and distance from (number of) health institutions.

CHAPTER ONE: INTRODUCTION

1.1. BACK GROUND OF THE STUDY

Tuberculosis (TB) remains one of the most important infectious diseases worldwide, with approximately one-third of the world's population infected with the Mycobacterium tuberculosis bacillus and more than 9 million new cases and 1.7 million deaths annually (WHO,2009). The current global epidemic of TB is enormous and growing and becoming more dangerous.

TB infection is caused by various strains of mycobacteria including Mycobacterium tuberculosis, Mycobacterium africanum and Mycobacterium bovis and is manifested as pulmonary or extra-pulmonary tuberculosis. About 85% of TB cases are pulmonary while 15% are extra-pulmonary tuberculosis. Extra pulmonary tuberculosis may involve the spinal cord, kidneys, skin, gastrointestinal tract, lymph nodes, genitourinary system and etc. However, pulmonary TB is much more important to deal with because it is the source of infection in the community.

Routes of transmission (TB) include the inhalation of air when a TB patient coughs, sneezes or spits and use of infected dairy products (Mycobacterium bovis). The probability of transmission per contact, per relevant unit time is general quite low. Mycobacterium may lodge in the person's lung and multiply, if the immune system in the lung is able to fight the bacteria and render it inactive (wall off) as a result of vaccination or treatment from TB, the person will develop latent TB which is not infectious and cannot harm others.

Infected individuals may remain asymptomatic over their entire live (Latent TB), but Active TB (the Clinical disease) can develop into pulmonary and extra pulmonary form. Extra pulmonary TB is common in children while pulmonary TB is frequent in adult. Mycobacterium tuberculosis, the casual agent of the disease is transmitted almost exclusively via pulmonary cases. Cases arising within Five (5) years from first infection are classified as primary tuberculosis while cases arising after five (5) years from first infection are known as secondary tuberculosis.

One third of the world's populations are thought to be infected with M. tuberculosis (about 2 billion people). However, every person infected with TB bacilli is not necessarily sick. Body's immune system fights the disease, but if the immune system is compromised due to old age, use of immune suppressive drugs or etc. about 5 to 10% of those with latent TB infection develop active tuberculosis at some point in their lives.

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 is to mean that, someone has chance of being affected by 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 many 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. As a consequence, knowledge about the distribution of TB plays a crucial role in formulating TB prevention and control program.

In most of the techniques used in traditional and molecular epidemiology studies, it is assumed that the person who contracts tuberculosis knew the infectious case, or at least has been in prolonged close contact with the case. Exploratory spatial data analysis is an approach consisting of a variety of statistical techniques intended to describe and visualize spatial distributions, identify atypical locations or spatial outliers, discover patterns of spatial association, clusters or hot spots (Longley, 1999) and suggest hypotheses without a necessary pre-conceived notion. By modeling the spatial nature of epidemiological data, it has been found that cases of disease tend to congregate at particular locations.

1.2. TUBERCULOSIS IN ETHIOPIA

TB is still a major cause of death worldwide, but the global epidemic is declining except sub Saharan Africa and Middle East countries. 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. Recent evidence demonstrates that TB prevalence (distribution) and TB death rates are globally decreasing after having reached a peak.

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.

Tuberculosis has been one the major causes of morbidity and mortality in Ethiopia for long. Accordingly, the Ethiopian Ministry of Health and its stakeholders have their unreserved and integrated efforts on this health problem. However, the direction to where tuberculosis in Ethiopia is heading hasn't been well analyzed and unpackaged by epidemiologically relevant factors. The incidence rate of tuberculosis is increasing in Ethiopia at a rate of 5 new tuberculosis cases per 100,000 populations per year. Urban agro-ecological zones have been more affected by the disease throughout the ten-year period. Extra-pulmonary rate and smear-negativity has shown a modest increment during the study period.

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.

Recent research paper (Habte, 2011) analysis the distribution of TB in North Shoa Zone of Oromia regional state, Ethiopia. The study revealed that incidence of TB disease is clustered to some woreda of the zone due to the difference in population density, prevalence of HIV cases, and topology of the woredas.

Due to the nature of its transmission and its significant effect on socio-economic aspect of society all over the world, recently researchers have been dealing on formulating new approaches, such as mapping the location with high cluster of cases with exploratory spatial analysis, for contributing to basic elements of TB control and therefore, this

study will attempts to identify and quantify location(areas) with high incidence of TB on one way and then to assess factors that govern its spatial distribution.

1.3. DISEASE MAPPING AND CONCEPT OF SPATIAL DEPENDENCE

Person, place, time: these are the basic elements of outbreak investigations and epidemiology. Historically, however, the focus in epidemiologic research has been on person and time, with little regard for the implications of place or space even though disease mapping has been done for over a hundred years. The development of geographic information systems (GISs) over the last 20 years has provided a more powerful and rapid ability to examine spatial patterns and processes. This, in turn, has fostered the discussion of such policy relevant issues as health services and planning (Matthews SA, 1990), as well as the use of GISs for epidemiologic investigations and disease surveillance.

The logic of using geography to study disease or health care is derived from appreciation of factors causing non-uniformity of disease distribution (Mayer JD). These factors include physical and environmental factors like: social, economic, cultural factors; and genetic factors. For example, diseases may be associated with environmental pollution, linked to individual or group behaviors, or associated with a genetic predisposition. In turn, all of these factors may have spatial distributions influencing the extent and intensity of a particular disease.

GISs, when combined with spatial analytical methods, may be helpful in the study of health care and health care delivery (Gesler W, 1986). Nearby locations are likely to possess similar attributes; or in other words, everything is related to everything else and near things are more related than distant things (Tobler W, 1979). These features of spatial data create needs for special analytical techniques so called spatial data analysis and it becomes familiar methods to be considered every time a project involving geography (i.e., location) is attempted.

Many epidemiological studies have been using traditionally Standardized Mortality Ratio (SMR) for mapping of disease incidence (mortality rates), however, the use of crude rates to estimate rare disease risks in small areas such as health units, census areas or administrative zones, is problematic since it does not account for the high variability of population sizes over the different regions, or the spatial patterns of the regions under study.

Alternatively, to improve the estimate of relative risk by employing smoothing tools on SMR, Spatial autoregressive models from classical statistics, and Conditional autoregressive model (CAR), Hierarchical model, and full Bayesian method from Bayesian methods are widely used, where all the methods of Bayesian methods proceeds through borrowing information from neighborhoods across the map.

1.4. STATEMENT OF THE PROBLEM

A bacillus, Mycobacterium tuberculosis, which is propagated through the air, cause tuberculosis. TB often will remain virtually asymptomatic for long periods of time, which serves to increase the temporal and spatial lapse between infection and diagnosis. Since

TB requires prolonged contact in order to spread, the disease tends to occur in spatial "clusters." The need for a spatial perspective on the tuberculosis epidemic is evident when examined in light of both the manner in which the disease is spread, and of potential mitigation strategies.

The uneven spatial distribution of the disease is a natural corollary of the nature of the disease itself. Tuberculosis is generally spread by close and prolonged contact with an infected individual, and the conditions, which lead to this sort of contact, are not randomly distributed (Kline.et. al, 1995).

An understanding of the spatial distribution patterns of tuberculosis is useful in targeting mitigation efforts. Resources for the combating of tuberculosis, while growing, are not infinite; understanding where occurrences of tuberculosis are greatest, or where outbreaks are most likely, allow the focusing of health - related resources on the areas where need is greatest, thus providing efficient treatment and mitigation with a minimum of public expenditure.

Based on this fact the main target of the study is to identify and quantify the spatial pattern of TB prevalence in East Hararghe Zone, including Harari Regional State which is engulfed by the zone. The analysis of the data is facilitated by geo-spatial software's like **Arc GIS 9.3, Geoda**, and **Win Bugs**.

1.5. OBJECTIVE OF THE STUDY

General objective

The main objective of the study is to identify and characterize the spatial distribution of TB cases in East Hararge Zone of Oromia Region including Harari Regional States, Ethiopia.

Specific objectives

- To explore spatial pattern of TB cases and identify high risk areas
- To assess the spatial dependence of TB prevalence on local environmental risk factors
- To fit spatial models for the TB distribution and examine significance of the model parameters

1.6. SIGNIFICANCE OF THE STUDY

The result of the study may help government and concerned bodies to formulate policy and implement prevention strategies based on pattern of TB distribution in the study area and also for identification of risk factors that are mostly related to incidence of TB cases.

1.7. LIMITATION OF THE STUDY

The study employed secondary data which was aggregated at woreda level; the number of TB cases that we encountered may be small portion of actual TB cases, especially in rural areas. On other way, the data analysis does not contain the spatial information of smaller groups of affected individuals. Consequently, the study assumes that every individual in each woreda have equal risk of exposure, therefore group level exposure represents individual level exposure.

1.8. THESIS ORGANISATION

This thesis comprises five chapters:

Chapter 1: Includes introduction, background of TB disease, statement of the problem, thesis objectives, limitation and significance of the study.

Chapter 2: focuses on review of the literature, reviewing the spatial methodology for TB prevalence. The applications of spatial related statistical models on disease mapping and testing the significance local risk factors are also discussed in this chapter.

Chapter 3: describes the study area, data sets, statistical tools and methods applied in the study.

Chapter 4: presents the result of the study and discussion of the results

Chapter 5: concludes the findings of the study and presents recommendations.

Ho (2004) reviewed the effect of socio cultural and politico-economic factors on distribution of tuberculosis based on a case study of Chinese immigrants in New York, United State. The objective of study was to incorporate the effect of socio cultural factors into the explanations for and management of tuberculosis and extent of the influence of cultural, environmental and politico-economic factors in contributing to the disease. To facilitate the study five groups of informants—public health workers, Chinatown biomedical doctors, Chinatown's practitioners of traditional Chinese medicine, Chinese laborers, and Chinese tuberculosis patients were included. The result of the study suggested that cultural, environmental, and politico-economic forces shaping tuberculosis and supports an emerging theorization of tuberculosis that encompasses a heterogeneous collection of factors.

Munch.et.al (2003) used the exploratory disease mapping techniques to identify spatial patterns of tuberculosis in South Africa. The study considered demographic factors like gender, sex, crowding, unemployment and alcohol use as risk factors in the transmission of tuberculosis. A dot map of TB cases implied that cases are concentrated in the southern and western parts of the country, and positive correlation was found between: the average caseload (all bacteriological-positive adults) and crowding, the average tuberculosis case load and unemployment, and also between crowding and unemployment. Poisson regression also confirms the correlation of crowding, unemployment and concentration of drinking places with tuberculosis cases.

Spatial Variations of Tuberculosis in the State of Kansas, United States for the period 1990 – 2000 was analyzed by Thomas and Richard (2004). A GIS raster grid was used in

order to find out high occurrence of tuberculosis respective of population size, and bivariate regression was used to identify what particular social, economic or residency parameters placed county residents at the highest risk. The finding of the study showed that there is a distinct spatial clustering in cases of tuberculosis in Kansas, largest number of cases in Sedgwick, Johnson, Wyandotte, and Finney counties. From regression analysis age and gender were found to have significantly risk factors than the general population average.

Nyirenda (2005) studied the spatial distribution of TB cases in Malawi for the period 1994-2004. The study identified distribution of diagnostic services and differences in health seeking behavior among different populations as contributing factors for spatial distribution of TB in different districts of Malawi. The result also revealed that women are more affected by TB at younger ages, while men are found to be more affected at older age. Parent – child relationship, alcohol and smoking and occupational exposure is found to be risk factors which are non spatial. The study suggests that for the incidence of TB in Malawi the strongest risk factors are Poverty, HIV infection, Household contact with index case, Overcrowding and age.

Friedrich (1998) performs spatial data analysis on TB distribution for identifying whether TB disease is independent of location or geographical factors. The study is undertaken with help of GIS and for spatial cluster detection Moran's I, Geary's *C* were employed. The result of the study indicated that a geographical data usually exhibit some amount of spatial dependency, local indicators of spatial autocorrelation depicts that there is a correlation between the values of TB case in neighboring districts.

Matthew (2005) analyzes the spatial distribution of TB incidence rate in Texas State, USA. For the accomplishment of the study total TB cases in Harris district of Texas during the period of 1995-1998 were examined to identify whether TB cases are spatially clustered or randomly distributed. The study employs spatial analytical techniques: mainly Global Moran's *I*, Geary's *C* and Local Indicators of spatial autocorrelation with GIS software. The result of the study shows that only two neighboring areas (districts) are found to be spatially correlated.

Venkatesan and Srinivasa (2010) conducted a study on modeling the spatial variogram of tuberculosis for Chennai district (which contains 28 wards) in India. Three different models of variogram namely spherical, exponential and Gaussian models were used in detecting the spatial dependence of tuberculosis in Chennai and then compared with respective theoretical variogram models fitted to tuberculosis cases. On the other hand the spatial mean, spatial standard deviation and spatial standard deviational ellipse were calculated. The result showed that spatial dependence exists between small distances of tuberculosis cases and spherical model is a better fit followed by Gaussian and exponential model.

William (2008) conducted a study on spatial and temporal-spatial clusters of tuberculosis in Ceara state, Brazil. Data on TB case for the period 1995-2006 in Ceara were employed by aggregating at various spatial scales such as state, municipality and with gender, urban/rural, age group, education as variable of interest. The study identified three hot spots in Fortaleza, Sobral, and Itapage municipalities, which are located on an east-west linear axis in the north of Ceara State while the entire southern region of Ceara was identified as a cold spot.

Venkatesan.et.al (2012) studied the spatial pattern of TB in India using the Bayesian conditional autoregressive model. They used National Family Health Survey data on tuberculosis to compare traditionally standardized Mortality ratio (SMR) and Conditional autoregressive (CAR) for smoothing the relative risk of TB case. Markov chain Monte Carlo (MCMC) simulation technique was used for disease mapping of CAR model and it was found to be the better model based on Deviance information criterion. The results revealed that northeastern of India states having higher risk of tuberculosis than other regions, and the study suggests that Bayesian CAR method is useful tool for modeling of tuberculosis.

Omoleke (2012) conducted a study to compare the incidence rate of tuberculosis in developing and developed world. For the study purpose Nigeria was taken from developing countries (Africa) and UK from developed. Result of the study reveals that the prevalence of HIV/TB is grossly higher in Nigeria than the UK (497: 15 per 100, 000 populations), the incidence is much higher in Nigeria (295/100,000 population) than the UK rate of 12/100,000 population. Mortality due to TB is grossly higher in Nigeria (67/100,000 population) than the UK with a much lower rate of 0.57/100,000 population. The empirical observations of the study suggest that surveillance is weak and ineffective in most African countries (compared to Europe) and socio-economic factors and political condition of the country was also found to be significant factor for prevention and control of TB.

Touray.et al. (2010) analyzes spatial distribution of TB incidence in Greater Banjul, of Gambia with help of spatial scan statistics to identify areas with high incidence rate. The finding of the study showed that out of all total TB cases registered 84% were permanent

residents with 88% living in 37 settlements, with two districts high and low incidence rates. Over dispersed Poisson regression and negative binomial regression models were fitted for TB distribution of study area, by identifying population density as significant explanatory variable.

Sudre.et.al (1996) studied risk factors for tuberculosis among HIV-infected patients in Switzerland. Univariate and multivariate logistic regression models were used to assess the association between the occurrence of tuberculosis and risk markers, such as: demographic characteristics of patients (age and gender); transmission category; country or region of origin; year of registration; and CD4+ cell count at the time of registration. The result revealed that patients from industrialized countries had a risk of tuberculosis similar to those from Switzerland whereas the risk among patients from Eastern Europe, Brazil and Africa was markedly higher. Age, sex and HIV-transmission category did not appear to increase the risk of tuberculosis after adjustment for other patient characteristics.

Abera.et.al (2009) conducted a study on assessing the association between HIV and TB in Oromia Regional State, Ethiopia. The study also identifies the spatial distribution of TB/HIV by forming cluster according to residential area (rural and the urban areas). Result of the analysis indicates that strong positive association was observed between TB and HIV using Spearman's correlation test and the incidence of TB and HIV was higher in urban areas or towns as compared to the rural areas.

Ismael (2008) analyzed the distribution of TB cases in rural and urban areas of Alamata woreda of Tigray Region, Ethiopia. For the study purpose, data of TB cases from two

urban and three rural kebeles of the woreda was considered. Regression analysis of TB cases with demographic factors is undertaken and correlation coefficient was calculated for each factors. The finding revealed that the distribution of TB was high in urban areas than in rural areas, and the Spearman's correlation coefficient analysis indicates no significant association between TB and any of the demographic variables considered.

Habte (2011) conducted a study on spatial distribution of TB cases in North Shoa Zone of Oromia Region, Ethiopia. TB cases (all type) of each woredas of the Zone which is recorded in one year (2008) is used for the analysis. Variables considered were Population density, HIV prevalence, and number of health centers in the woredas. Moran's I and Geary's C statistics revealed significant cluster of TB cases in six woredas of Zone, dissimilar value in only one woreda. Over dispersed Poisson regression model and Negative Binomial regression was fitted, were both models identified strong association of TB load to the number of Health centers, HIV prevalence, and Population density. The study also suggested that when the count data is over dispersed, Negative binomial regression model is preferred.

CHAPTER THREE: DATA AND METHODOLOGY

3.1. SOURCE OF THE DATA

The data that we used for studying spatial pattern of TB cases are the secondary data of total TB counts (2004, E.C), compiled and aggregated at woreda level by Zonal Health Bureaus. Data on TB load includes all type of TB cases which are specified and recorded in both governmental and private health institutions. The health institutions are; Hospitals, Health centers, Clinics, Health posts, and Drug shop.

3.2. STUDY AREA AND VARIBLES OF THE STUDY

The study area; East Hararge Zone, is found in Eastern part of Oromia region, Ethiopia. Administratively the Zone is subdivided in to 18 woredas namely Babile, Bedeno, Chinaksan, Deder, Fadis, Gola-oda, Goro-gutu, Girawa, Gursum, Haromaya, Jarso, Kersa, Kombolcha, Kurfa-chele, Kumbi, Melka-belo, Mayu-muleke, and Meta. Geographically it is located $7^{0}32^{1}$ - $9^{0}44^{1}$ North latitude and 41^{0} 10^{1}-$43^{0}16^{1}$ East longitudes, and shares boundaries with West Hararge Zone from the west, Bale Zone from the south, Somali regional state from the east and southeast, and Dire-Dawa Administrative Council from the North.

Harar town which serve as the capital city for both East Hararge zone and Harari Regional State is at a distance of 526 kilometers from Addis Ababa to the East side of the country. With a total land coverage of 25,747.91 km, ^{2} and a total population of **2,917,502,** the study area has a density of 113 people per square km (2004, E.C). Altitude of the zone ranges from 500 to 3405 meters above sea level, and the three major climatic categories of the zone are temperate tropical high lands that constitutes 11.4% of

the area, Semi temperate (tropical rainy mid lands) accounts for 26.4% and the rest (62.2%)of the total area is Semi arid (tropical dry or arid).

Harari Regional State is one of the Regions (States) in Ethiopia located in Eastern part of the country, surrounded by woredas of East Hararghe Zone in all directions. The region has no administrative zones or woredas, and has a total population of **183,415**, of which majority are living in urban areas. The regional state has an estimated area of 371 square-km; estimated density of 494 people per square-km in 2004 E.C.

Variables of the study

The dependent (response) variable of the study is total TB cases, comprises all forms of TB case.

The independent (explanatory) variables considred are:

i. Population density: the ratio of total population size to the total land coverage, given in per square km.

ii. HIV prevalence. The ratio of HIV population to the population at risk multiplied by 1000.

iii. Total number of health centers_ governmental and nongovernmental institutions that give health services for the society; Number of hospitals, health stations, health posts, clinics, and Drug shops.

3.3. METHODOLOGIES OF THE STUDY

3.3.1. SPATIAL AUTOCORRELATION

Spatial autocorrelation is quantified in spatial analysis through the use of spatial statistics. Spatial statistics are used to detect patterns of spatial autocorrelation that represent areas of either high or low disease risk (Waller & Gotway, 2004). These patterns, which often represent areas of significant existence or non existence of the disease, are referred to as clusters. Many spatial statistics that detect clusters also describe cluster morphology, which can be the geographic size

Geographical data are correlated in space that is data in close geographical proximity is more likely to be influenced by similar factors and thus affected in a similar way.

Deeply rooted in the notion that geographic location matters, one testable assumption is that near things are more related than distant things; a concept often referred to as Tobler's first law of geography, "Everything is related to everything else, but closer things more so". The spread of infectious disease is closely associated with the concepts of spatial and spatiotemporal proximity, as individuals who are linked in a spatial and a temporal sense are at higher risk of getting infected (Dirk, 2008).

Thus the knowledge of spatial and temporal variation of disease and characterizing its spatial structure is essential for understanding population interaction with its environment. In the case of tuberculosis, spatial correlations are present at both short and large scales, reflecting the transmission of tuberculosis infection and the effects of environmental factors (Venkatesan and Srinivasan, 2008).

Cluster detection within a spatial analysis can be undertaken using a variety of spatial statistical tests, many of them characterized as global or local, where both global and local statistics are used to identify areas of clustering in studies that do not have a pre-determined hypothesis about where clusters may be located.

The common principle of the different test statistics for spatial autocorrelation is that the comparison of the value of the statistic for a particular data set to its distribution under the null hypothesis of "no spatial autocorrelation." Such a null hypothesis implies that space does not matter, or, in other words, that the assignment of values to particular locations is irrelevant. Hence, it is only the values that provide information to the analyst, and "where" they occur does not add any insight. In contrast, under the alternative hypothesis of spatial autocorrelation (spatial dependence, spatial association), the interest focuses on instances where large values are systematically surrounded by other large values, or where small values are surrounded by other small values referred to as positive autocorrelation or where large values are surrounded by small values (and vice versa) which is called negative spatial autocorrelation (Cressie, 1993).

General tests are carried out with what are called "global'' statistics; again, a "test for the detection of clustering". Here there is no a priori idea of where the clusters may be; the methods are aimed at searching the data and uncovering the size and location of any possible clusters. Single summary value characterizes any deviation from a random pattern. On the other hand, "Local" statistics are used to evaluate whether clustering occurs around particular points, and hence are employed for both focused tests and tests for the detection of clustering. Local statistics have been used in both a confirmatory

manner, to test hypotheses, and in an exploratory manner, where the intent is more to suggest, rather than confirm, and hypotheses.

The important assumption to determine the distribution of a test for spatial autocorrelation is that data follow an uncorrelated normal distribution. Based on the properties of this distribution, the moments of the statistic under the null hypothesis can be derived analytically; the other approach is non-parametric and exploits the interpretation of the null hypothesis as being non-spatial.

In other words, each observation can be assumed to occur with equal probability at all locations and the approach is referred to as the randomization assumption.

3.3.2. SPATIAL WEIGHTS AND NEIGHBORHOODS

Spatial autocorrelation measures require a weight matrix that defines a local neighborhood around each geographical area or units. The value at each areal unit is compared with the weighted average value of its neighborhood. This weight matrix is a square symmetric R x R matrix with (i,j) elements equal to 1 if regions i and j are neighbors of one another (or more generally, are spatially related), and zero otherwise. By convention, the diagonal elements of this "spatial neighbors" matrix are set to zero. Spatial autocorrelation measures such as Moran's I require a weights matrix that defines a local neighborhood around each geographic unit. Weights can be constructed based on contiguity to the polygon boundary (shape) files, or calculated from the distance between points (points in a point shape file or centroids of polygons).

The formula for each weight is

$$wij = \frac{C_{ij}}{\sum_{i}^{N} C_{ij}}$$, with C_{ij}=1 if i and j are linked, and C_{ij}= 0 otherwise.

Generally there are three kinds of weight matrices: contiguity, distance and K-nearest neighborhood.

a. Contiguity weight files

Most analyses of spatial autocorrelation adhere to a common definition of contiguity. Contiguity refers to what polygons are selected as neighbors for a single target polygon. Under contiguity classification we have Rook, Bishop and their combination (Queen Contiguity).

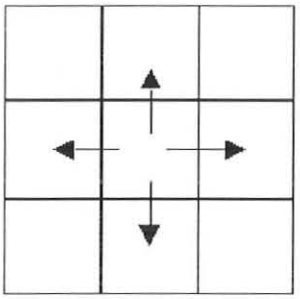
Rook contiguity (so called after the movement of the chess piece): two regions are neighbors if they share (part of) a common border (on any side).

Bishop continuity: two regions are spatial neighbors if they meet at a "point". This is the spatial analog of two elements of a graph meeting at a vertex.

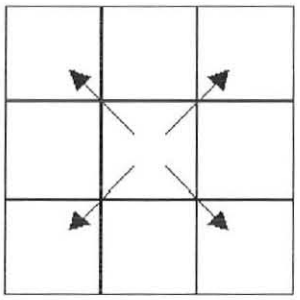
Queen contiguity: this is the union of Rook and Bishop Contiguity. Two regions are neighbors in this sense if they share any part of a common border, no matter how short.

The description of contiguity definition for neighborhood is given graphical:

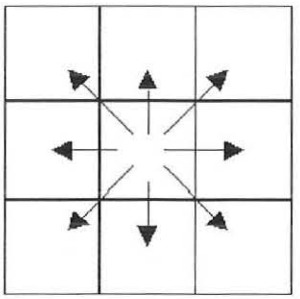
Rooks Case



Bishops Case



Queen's (Kings) Case



b. Distance weights

Under this category XY-coordinates are used to automatically calculate distance between points or centroids of polygons and specify the cut-off point (threshold distance) to determine the minimum distance for two units to be considered as neighbors.

c. K-Nearest neighbors

First the exact number of neighbors that a unit should have is specified and then appropriate software like Geoda is used to find areas that are nearest to the unit.

3.3.3. GLOBAL MEASURES OF SPATIAL AUTOCORRELATION

The literal meaning of spatial autocorrelation is self-correlation (autocorrelation) of observed values of a single attribute, according to the geographical (spatial) ordering of the values. The most common techniques for measuring Global spatial autocorrelation are Moran's Index statistic (Moran, 1948) and Geary's C statistic (Geary, 1954). These tests indicate the degree of spatial association as reflected in the data set as a whole. They both necessitate the choice of a spatial weights matrix. Moran's Index is based on cross products to measure value association, while Geary's C employs squared differences.

i. Global Moran's I

It provides a single measure of spatial autocorrelation for an attribute in a region as a whole. Formally, Moran's I for N units of variable y, is expressed as:

$$I = \left(\frac{N}{W_o}\right)\frac{\sum\sum w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum(y_i - \bar{y})^2} \quad \text{or} \quad I = \frac{N}{\sum\sum w_{ij}}\frac{\sum\sum Z_i W_{ij} Z_j}{\sum Z_i^2} \qquad \text{.....(1)}$$

Where I is Moran's I statistic, N is the total number of units (woredas), y is observed values of attributes (TB) cases at each location (woreda), w is weight matrix (connectivity) which contains information of the location, Z is standardized values of response variable (TB) cases, and W_o is the normalizing factor equal to the sum of the elements of the weight matrix ($W_o = \sum\sum w_{ij}$, i≠ j).

Under the normal and randomization assumptions, the resulting standardized-values are compared to a table of standard normal to assess spatial dependence. The null hypothesis (no spatial autocorrelation), will be rejected if the calculated value of $|Z_I| \geq Z_{\frac{\alpha}{2}}$, where

$$Z_{(I)} = \frac{I - E(I)}{\sqrt{\text{var}(I)}} \qquad \text{.....(2)}$$

$$E(I)_N = \frac{-1}{N-1} = E(I)_R \qquad \text{.....(3)}$$

The variance of the Moran's I vary with assumption of the normality and randomization:

Under normality assumption,

$$Var(I)_N = \frac{N^2(N-1)W_1 - N(N-1)W_2 - 2W_o^2}{(N+1)(N-1)W_o^2} \quad \text{.....(4)}$$

$$W_o = \sum_{i \neq j}^{N} w_{ij}, i \neq j, W_1 = \frac{1}{2}\sum\sum(wi_j + W_{ji})^2, W_2 = \sum_i [\sum_i wi_j + \sum_j w_{ji}]^2$$

and with randomization assumption,

$$Var(I)r = \frac{N(W_1(N^2 - 3N + 3) - NW_2 + 3W^2{}_o)}{(N-1)(N-2)(N-3)W_o^2} - \frac{K(W_1(N^2 - N) - 2NW_2 + 6W^2{}_o)}{(N-1)(N-2)(N-3)W_o^2} - \left(\frac{1}{n-1}\right)^2 \quad \text{.....(5)}$$

all notation are as in (4) and

$$K = \frac{N\sum_{i=1}^{N}(y_i - \bar{y})^4}{\sum_{j=1}^{N}((yi - \bar{y})^2)^2}$$

ii. Global Geary's C

Geary's C statistic is given by taking cross product of deviation of each location from its neighbor than from its average value. Algebraically according to Geary (1954),

$$C = \left(\frac{N-1}{2W_o}\right)\frac{\sum_i\sum_j w_{ij}(y_i - y_j)^2}{\sum_i (yi - \bar{y})^2} \quad \text{.....(6)}$$

Like Global Moran's Index, the standardized values of the observation will be calculated and the null hypothesis of no spatial autocorrelation will be rejected if the calculated value, $|Z_c| \geq Z_{\frac{\alpha}{2}}$,

$$Z_{(c)} = \frac{C - E(C)}{\sqrt{\text{var}(C)}} \quad \text{.....(7)}$$

$$E(C)_N = E(C)_R = 1 \quad \text{.....(8)}$$

Equation (7) and (8) given above represents the standardized and expected values respectively

The variance of Geary's C also varies with assumption of normality and randomization:

$$Var(C)_N = \frac{(2W_1 + W_2)(N-1) - 4W_o^{\,2}}{2(N+1)W_o} \quad \text{.....(9)}$$

$$Var(C)_R = \frac{W_1(N-1)(N^2 - 3N + 3 - K(N-1))}{Wo(N(N-2)(N-3)} + \frac{(N^2 - 3 - K(N-1)^2)}{N(N-2)(N-3)} - \frac{(N-1)W_2(N^2 + 3N - 6 - K(N^2 - N + 2)}{4N(N-2)(N-3)Wo^2} \quad \text{.....(10)}$$

The notation used is the same as in (5).

Interpretation: A value of Moran's I that is larger than its theoretical mean of –1/N-1, or, equivalently, a positive z-value, points to positive spatial autocorrelation. In contrast, for Geary's C, positive spatial autocorrelation is indicated by a value smaller than its mean of 1, or by a negative z-value.

Note: the assumption of normality in spatial sense is the same as free sampling (sampling with replacement) where as randomization is like sampling without replacement. i.e. in Free (or normality) sampling assumes that the probability of a polygon having a

particular value is not affected by the number or arrangement of the polygons, analogous to sampling with replacement where as non-free (or randomization) sampling assumes that the probability of a polygon having a particular value is affected by the number or arrangement of the polygons (or points), usually because there is only a fixed number of polygons.

3.3.4. LOCAL MEASURES OF SPATIAL AUTOCORRELATION

Indicators of global spatial autocorrelation either Moran's I or Geary's C, though they can identify whether or not clustering is occurring, cannot specify the location of clusters or how spatial dependency can vary from one place to another. Local spatial statistics are used to quantify clustering within smaller areas of a larger study area, and in many instances can be seen as smaller partitions of the global spatial statistical analysis (Fotheringham.et.al, 2002). More generally, local indicator of spatial autocorrelation (LISA) shows statistically significant groupings of neighbors with high and low values around each region in the study area and therefore, aim of using this LISA is to identify 'hot spots' or cold spots (Anselin, 1995).

For this study local Moran's I and Local Ord and Getis G* statistics have been used for identification of cluster (hot spots) at local level.

i. Local Moran's I

Local Moran's I for each observation measures the extent of significant spatial clustering of similar values around that observation, and algebraically it is given as

$$I_i = \frac{\sum w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{(y_i - \bar{y})^2}, \quad \text{Or} \quad I_i = Z_i \sum w_{ij} Z_j \quad \ldots\ldots\ldots\ldots\ldots\ldots (11)$$

Z_i and Z_j are standardized values of attribute (TB cases) for units i and j, where j is among the identified neighbors of i according to the weight matrix and all other notation are in (5).

Analogous to Global counterpart, the null hypotheses which states no spatial clustering will be rejected if the computed Value, $|Z_{Ii}| \geq Z_{\frac{\alpha}{2}}$, algebraically it is given by

$$Z_{Ii} = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}} \quad \text{.....(12)}$$

$$E(I_i) = \frac{\sum_j w_{ij}}{N-1} \quad \text{.....(13)}$$

$$Var(I_i) = \frac{W_i(N-b_2)}{N-1} + \frac{2W_j(2b_2-N)}{(N-1)(N-2)} - \frac{W_i^2}{(N-1)^2} \quad \text{.....(14)} \ ,$$

$$b_2 = \frac{M4}{M2^2}, \qquad M4 = \frac{\sum_i y_i^4}{N}, \ M2 = \frac{\sum_j y_j^2}{N}, \ W_i = \sum_{j \neq i}^{N} w_{ij}^2 \qquad Wj = \sum_{h \neq i}^{N} \sum_{k \neq i}^{N} w_{ik} w_{ih}$$

Note: the subscript i,j above shows the location(places) that are used in the computation, and all other notation are the same in (5). The interpretation for the local Moran's I is the same as that of global counterpart

ii. Local Ord and Getis, G*.

The Ord and Getis **Gi*** local statistics for measuring spatial autocorrelation (Ord and Getis, 1995) is given as follows:

$$Gi^* = \frac{\sum_j w_{ij} y_j - (\sum_j w_{ij} + w_{ii})\bar{y}}{S \times \left[\frac{NS_i^* - (\sum w_{ij} + w_{ji})^2}{(N-1)} \right]^{\frac{1}{2}}} \quad \text{.....(15)}$$

$$E(Gi^*) = \frac{\sum_{j=1}^{N} w_{ij}}{N} \;,\; S^*_i = \sum (wij^2 + wii^2) \text{ and } \operatorname{var}(Gi^*) = \frac{W_i^*(N - E(G_I^*))Y_i^*}{N^2(N-1)Y_k^*} \quad \text{where}$$

$$W_i^* = \sum_j w_{ij} \;,\; yi^* = \frac{\sum_j y_j}{N} \;,\; y_k^* = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(y_i y_j)^2}{N} - y_i^* \text{ and } w_{ii} \text{ is a weight in the case in}$$

which i is in its own neighborhood set, $\bar{y}$ is average of the dependent variable (TB Cases). The formula given in (15) shows what portion of the total sum of all values is represented by the values at, and near, locations i (Getis and Ord, 1992).

Interpretation: Positive values of the **Gi** * statistic indicate that high values are spatially clustered with other high values of the random variable. Negative values of **Gi** * statistic indicate that low values are spatially clustered with other low values of the dependent variable.

Note that a consequence of this is that the **Gi** * statistic, unlike Moran's I, cannot distinguish cases of positive spatial autocorrelation from cases of negative spatial autocorrelation (Getis and Ord, 1992).

3.3.5. Testing for spatial dependence in regression residuals

The error term in a regression can be considered to contain all ignored elements. If some of these show a significant spatial pattern, it will be reflected in a spatial pattern for the error terms and dependent variable.

There are two most common types of diagnostics available to test for spatial dependence in the residuals of a linear regression: an extension of Moran's I statistic to regression residuals (Cliff and Ord, 1972) and a simple test based on the Lagrange Multiplier (LM) principle (Burridge, 1980). Both tests; test for residual autocorrelations (Moran's I) and LM are derived from the results of an ordinary least squares estimation in a standard regression model.

Using standard matrix notation the spatial dependence in the error term is specified as:

$$y = X\beta + \varepsilon$$
$$with \varepsilon = \lambda W\varepsilon + \xi \quad \text{.....(16)}$$

Where y is N by 1 vector of observations on the dependent variable, X is N by k matrix of observations on explanatory variables with matching regression coefficients in the $k \times 1$ vector β and ε is a N by 1 vector of error terms. ε is assumed to follow autoregressive with coefficient λ and a white noise error ξ

i. Moran's I diagnostic for spatial error dependence

Moran's I statistics; test for error spatial autocorrelation from regression residuals (Cliff and Ord, 1972) is given by:

$$I = \frac{N}{Wo}\frac{e'We}{e'e} \quad \text{.....(17)}$$

e is an N by 1 vector of residuals, W is the spatial weights matrix and W_0 is the normalizing factor (the sum of all weights) and N is number of observations (woredas). The Moran test statistic I (under Ho: $\lambda = 0$) is distributed as F (F-distribution) with N and Wo degree of freedom.

$$\frac{e'we/S}{e'e/\sigma^2} \sim \frac{\chi 2/W_o}{\chi 2/N} \sim F\alpha(W_o, N) \quad \text{.....(18)}$$

Test decision: the null hypothesis (Ho) will be rejected if the computed value is greater than tabulated value.

ii. Lagrange Multiplier test

The Lagrange Multiplier (LM) or score test for spatial error autocorrelation is very similar in form to Moran's except normalizing factor. But in function, Moran's I is used only to detect the existence of spatial autocorrelation, while LM test has additional use for choosing between a spatial error and a spatial lag alternative (Anselin, 1999), and therefore can be used for the inference of spatial autoregressive coefficients.

The expression for the LM test against a spatial error alternative (SEM) was original given by Burrigde (1980) to test the null hypothesis of no spatial error autocorrelation, $H_o : \lambda = 0$, as:

$$LMer = \frac{\left(\frac{e'we}{\sigma^2}\right)}{tr[(w+w')w]} \quad \text{.....(19)}$$

Here, tr stands for the matrix trace operation (sum of the all elements of the diagonal), and σ^2 is replaced by its estimate, $\frac{e'e}{N}$. The LM statistic is asymptotically distributed as a χ^2 variate with one degree of freedom.

Similarly, the LM test against a spatial lag alternative (SLM) was outlined to test the null hypothesis of no spatial dependence due to the response variable, or ($H_o : \rho = o$), where ρ represents the autoregressive coefficients for spatial lag.

The Lagrange multiplier test for spatial lag is given by:

$$LM_{lag} = \frac{\left(\frac{e'we}{\sigma^2}\right)}{D} \quad \text{.....(20)}$$

$D = (WX\beta)'(I - X(X'X)^{-1}X'(WX\beta)/\sigma^2 + tr[(W + W')W]$, all other notation is the same as in equation (17). This LM $_{lag}$ has also an asymptotic $\chi^2(1)$ distribution.

Note: If the test results for both LM diagnostic cannot reject the null hypothesis, it means that OLS is sufficient for modeling the distribution with specified spatial dependence.

iii. **Robust Lagrange Multiplier Diagnostics**

Robust LM test are test statistics used to support the test result given by its non robust test.

The robust LM for lag which is used to test the null hypothesis of spatial lagged dependent variable ($H_o : \rho = o$), is given by:

$$RLM_{lag} = \frac{\left(\frac{e'Wy - e'We}{\frac{e'e}{N}}\right)}{tr[(W'+W)W] + (WX\beta)'(I - X(X'X)^{-1}X'(WX\beta/S^2)^{-1}} \quad \text{.....(20)}$$

The robust LM for error which test against the null hypothesis of no spatial error dependence

$(H_o : \lambda = o)$, is given by:

$$RLM_{error} = \frac{\left[\left(\frac{e'We}{S^2} - T\right) - T*(e'WY/\frac{e'e}{N})\right]}{T - T^2(T*)} \quad \text{.....(21)}$$

$$T = tr[(W + W')W], and T* = (T + (WX\beta)'(I\_X'X)^{-1}X'(WX\beta)/S^2)^{-1}$$

The interpretation and the rejection rule for both test statistics is the same their non robust counterpart.

3.4. MODELING SPATIAL DEPENDENCE

Once the distribution is identified, another resolution is enhancement in spatial data that offers an opportunity to model disease, which helps us in identifying factors that affect TB distributions.

Spatial analysis focuses on counts from small areas with relatively at risk and few cases expected during the study period, such instance require models appropriate for count or rare outcome (Manfred.et.al, 2011)

Modeling spatial interaction that arise in spatially referenced data is commonly done by incorporating the spatial dependence in to the covariance structure either explicitly or implicitly via an autoregressive model. In case of lattice or areal data which can be

defined as observation associated with a fixed number of areal units such as counties, districts and census zones, the two most common Auto regressive model used are Bayesian conditional autoregressive model (CAR) and the spatial autoregressive model (SAR).

Both of these models produce spatial dependence in the covariance structure as a function of neighborhood matrix, W and often a fixed unknown spatial correlation parameter (Wall, 2004). The description of each of the models is given below one by one with their respective estimation and diagnostic techniques.

3.4.1. SPATIAL AUTOREGRESSIVE MODELS

The treatment of spatial data analysis from the lattice data perspective focuses on two main issues: testing for the presence of spatial association, and the estimation of regression models that incorporate spatial effects. The indication of a significant pattern of spatial clustering given by a test for spatial autocorrelation is only an initial step in the analysis of spatial data.

Such an indication shows that the observations are more clustered than they would be under a random assignment, but it does not explain why such clustering occurs, nor which factors determine its shape and strength. In other words, the alternative hypothesis of "spatial autocorrelation" is too vague to be very useful in the construction of theory.

A concept that formalizes the way in which the spatial association is generated, is that of a spatial process, or spatial stochastic process. Roughly speaking, such a process expresses how observations (error term) at each location depend on values at neighboring

locations, i.e., spatial lag model and spatial error model. An alternative to the pure spatial process specification is the inclusion of "exogenous" explanatory variables in addition to the spatial effects.

Spatial autoregressive model is an autoregressive model that accounts for spatial effects in estimating model parameters. There are two common ways of incorporating spatial dependence, as an additional lag of dependent variables (spatial Lag model) and as lagged component of error term (spatial error model). Spatial lag model (SLM) consider the average weighted effect of its neighbor as other factor to determine the model (Anselin, 1988), and this weighted average is called the spatial autoregressive parameter. Spatial lag model, therefore suggests that the spatial autoregressive or the effect of spatial dependence exists only in dependent variable, but not due to other explanatory or error term.

In its simplest form, with only one order of contiguity used for the spatial lag terms, SLM can be expressed as:-

$$y = \rho Wy + X\beta + \varepsilon \quad \text{.....(22)}$$

Where y is a N by 1 vector of observations on the dependent variable, Wy is the corresponding spatial lag for weights matrix W (which control spatial autoregressive), ρ is a spatial autoregressive parameter and ε is a N by 1 vector of error terms and X is a N by k matrix with observations on the k exogenous explanatory variables with matching k by 1 coefficient vector of β. Spatial autoregressive (lag) model given in (equation, 22) can be interpreted in two ways: In a first perspective, the interest is finding how the

variable y relates to its value in surrounding locations (the spatial lag), while controlling for the influence of other explanatory variables. This is the case when the spatial pattern in a dependent variable is actually due to the spatial pattern of other variables with which it is strongly correlated.

A second interpretation is considered when the interest is really in the relation between the explanatory variables X and the dependent variable Y, after the spatial effect has been controlled for. This is often referred to as spatial filtering or spatial screening (Getis, 1990), and can be formally expressed as:

$$(y - \rho Wy) = X\beta + \varepsilon$$
$$(I - \rho W)y = X\beta + \varepsilon$$
$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon \quad \text{.....(23)}$$

Here, the matrix $(I - \rho W)$ is referred to as a spatial filter, with I as an identity matrix of dimension N by N.

A spatial error model (SEM) on other hand is a weighted average of the individual residuals of the neighboring locations (Anselin, 1992), incorporated to the model as an additional explanatory variable. The mathematical model for SEM is given by:

$$y = X\beta + \varepsilon$$
$$with \varepsilon = \lambda W\varepsilon + \xi \quad \text{.....(24)}$$

all notation used are the same the as in equation (16)

Hence, SEM postulates that the spatial influence on the dependent variable is coming through the error term, i.e. it assumes that spatial effect occurred in error term rather than with dependent or other explanatory variables of the model.

3.4.2. MAXIMUM LIKELIHOOD ESTIMATION

Due to multidirectional nature of spaces, ordinary least squares estimates (OLS) are biased, inconsistent and in efficient for estimating model parameters in spatial data analysis, and alternatively, Ord (1975) gives the maximum likelihood estimation methods for estimating the spatial lag and spatial error models.

Although MLE has the best efficiency among all available estimators when the assumption of error distribution is met, the computation is not as that of classic regression models. That is due to the two-directional nature of the spatial dependence, log likelihood results in a Jacobean term, determinant of a full $n \times n$ matrix, rather than sum of the log likelihoods associated with the individual observations.

i. Maximum likelihood estimate for spatial lag model

Assuming $\varepsilon \sim N(0, \sigma^2)$ the log likelihood function of spatial lag model is (Anselin, 1999)

$$\ln L(\sigma, \beta, \rho) = \ln|I - \rho W| - \frac{n}{2}\ln(2\pi) - \frac{-n}{2}\ln(\sigma^2) - \frac{(y - \rho Wy - X\beta)'(y - \rho Wy - X\beta)}{2\sigma^2} \quad \text{.....(25)}$$

Likelihood, (25) is non linear function of parameters and it is value that is to be maximized, Anselin (1988, ch. 12) suggests a way to do the estimation, by focusing first on β estimation as follows:

$$\begin{aligned} b &= (X'X)^{-1}X'AY \\ &= (X'X)^{-1}X'Y - \rho(X'X)^{-1}X'WY \quad \text{.....(26)} \\ &= bo - \rho bl \end{aligned}$$

method they took an empirical Bayes approach that shrinks the SMR's towards a local or global mean, where the amount of shrinkage may depend on local (spatial) and global variability.

Bayesian estimators are widely used in order to obtain reliable estimate for the relative risk when there are sub-areas with small population, i.e. if in the sub-areas we observe too few cases, then the estimated relative risk is very unstable and can lead to wrong conclusions. When a population exposed to risk is small the number of observed cases of a disease can be affected by a large variability due only to chance, and not due to environmental variables or particular behavior of the target population. The need to identify extreme rates for areas with small population or rare diseases led to the development of Bayesian estimation in disease mapping.

Bayesian Conditional Autoregressive (CAR) model is a disease mapping technique, which is used for smoothening of relative risk (Clayton, 1987). This model (CAR) provides some shrinkage and spatial smoothing of the raw relative risk estimate, and gives a more stable estimate of the pattern of underlying risk of disease than that provided by the raw estimates.

CAR borrows information from neighboring areas than from areas far away and smoothing local rates toward local (neighboring values). This method reduces the variance in the associated estimates and allows for the spatial effect of regional differences in whole populations.

The Bayesian (CAR) approaches of disease mapping combine two types of information, the information provided by the observed number of cases in each site described by the Poisson likelihood, $L(\theta/y)$ and prior information on the relative risks specifying their

variability in the overall map, summarized by their prior distribution (Mollie,1999). Supposing that the relative risks are correlated where the correlation is dependent on geographical proximity and that the relative risk can be considered to be Gaussian, estimating the relative risk using CAR gives (Besag, 1974 and Ord, 1975)

$$E(\beta_i / \beta_{j,j\neq i}) = \mu_i + \rho\sum W_{ij}(\beta_j - \mu_j) \quad \text{.....(32)} \quad \text{and}$$

$$Var(\beta_i / \beta_{j,j\neq i}) = \sigma^2 \quad \text{.....(33)}$$

Where W is weight matrix of the map, ρ spatial dependent parameter.

Yasni.et.al (2000) with the same assumption of Gaussian modified the mean and variance of the CAR model as following:

$$E(\beta_i / \beta_{j,j\neq i}) = \alpha + \frac{\rho\sum_{i\neq j} W_{ij}(\beta_j - \alpha)}{\sum_{j\in\partial i} W_{ij}} \quad \text{.....(34)} \quad \text{and} \quad V(\beta_i / \beta_{j,j\neq i}) = \frac{\sigma^2}{\sum_{j\in\partial i} W_{ij}},$$

∂ -is set of neighborhood location (site) for i and j.

Conditional autoregressive (CAR) models are commonly used to represent spatial correlation in lattice data, which arise in a wide variety of applications including education, epidemiology, agriculture, etc. The specification for CAR model relies on the conditional distribution of the spatial error terms. In this case, the distribution of e_i conditioning on e_{-i} (the vector of all random error terms minus e_i itself), i.e. instead of the whole e−i vector, only the neighbors of area i, defined in a chosen way, are use used.

The set of areal (lattice) units can form a regular lattice that exhibit spatial correlation, with observations from areal units close together tending to have similar values. A proportion of this spatial correlation may be modeled by including known covariate risk factors in a regression model (CAR).

This study employed two different models for describing spatial pattern of TB cases in study area; spatial autoregressive model (see, section 3.4) and Poisson-CAR model. For fitting CAR model, we used data of TB cases of each woredas, expected TB cases, and spatial feature of the study area partitioned in to woredas. Accordingly, Y_i (TB cases) occurring in area Ai is recorded, where the set of areas {Ai},i = 1, 2,…,n represents a partition of the region(Zone) under study. For each area Ai, the expected number of cases (Ei) is computed using reference rates (overall incidence risk of zone multiplying by population at risk for each woredas).

The distribution of the observed value of TB case (counts), Y_i is typically assumed to come from a Poisson distribution, as the diseases (TB cases) usually considered as rare and this distribution gives a good approximation to the underlying binomial distribution that would hold for each risk areas.

Assuming the Poisson distribution for the total number of TB cases, and employing Markov assumption for the spatial context; the property that the conditional distribution of each site, p (xi |x−i) depends only on a few components of x−i, called the neighbors of site i, we can apply CAR model that can be expressed as:

$$Y_i \sim \ poisson(\eta(i)) \quad \text{.....(35)}$$

$$\log(\eta(i)) = \log(E_i) + \alpha_o + \alpha_1 X_1 + \beta X_2 + \gamma X_3 + bi \quad \text{.....(36)}$$

Here $\mathbf{Y}_i$ is the observed count of the TB case (disease) in the i th region and E_i is the expected count within the same region. Y_i is assumed to be conditional independent given b_i, and then the likelihood for TB case is obtained as product of joint densities.

The parameter of interest (η (i)) is the relative risk that quantifies whether the area i has a higher risk or lower occurrence of cases than that expected from the reference rates, αo - Intercept term representing the baseline log relative risk of disease across the study region, $X_{i,\ i=1,2,3}$ are the covariates (X_1-population density,X_2-number of health centers, and X_3-HIV prevalence) in district i, with associated regression coefficients α, β, γ respectively. b_i is an area-specific random effect capturing the residual or unexplained (log) relative risk of disease in area i.

We often think of b_i as representing the effect of latent (unobserved) risk factors, to allow for spatial dependence between the random effects b_i in nearby areas, we may assume a CAR prior for these terms.

3.5.1. CAR Model and Prior specification

The model consists of three components: the likelihood for TB cases (count), the latent process model for log relative disease risk and the prior specification. Win BUGS (Bayesian Analysis Using Gibbs Sampling); package that has been designed to carry out Markov chain Monte Carlo (MCMC) computations for a wide variety of Bayesian models was used for analysing our data.

To fit the model in Win BUGS; the observed TB case, expected TB case, and covariates including population density, number of health center and HIV prevalence with adjacency matrix of study area were included for analysis. The prior for this model (mean and variance of random effect) is hyper prior, Gamma prior which is distributed with a small precision, thus taking a larger neighborhood structure into account. Results are based on Markov Field Monte Carlo (MCMC) simulation and an inverse distance-matrix

3.6. MODEL SELECTION METHODS

There are two common ways of comparing spatial autoregressive models with standard regression analysis. The first and the most simple one is to compare the outcome of diagnosing spatial dependence which is obtained by regressing TB case on only explanatory variables (OLS) by adding the weight matrix (section 3.3.5). the second means of selecting appropriate autoregressive model in classical statistics is to compare the values of Likelihood ratio test (LRT), Akaike information criterion (AIC) and Schwarz Information Criterion (BIC) Likelihood ratio test (LRT).

When the models are estimated by MLE, a likelihood ratio test (LRT) can be used to test whether or not spatial autoregressive models (SAR) make a significant improvement in model fitting over OLS, or, in other words, whether or not the improvement warrants the additional computation needed for SAR (Haining 1990, p. 142–145):

$$LRT = 2(\ln L_1 - \ln L_2) \quad \text{.....(37)}$$

L_1 and L_2 are the likelihoods of two models.

The null hypothesis is H_0, $\rho=0$; and the alternative hypothesis is H_1; $\rho \neq 0$. This LRT statistic approximately follows a $\chi 2$ distribution with degrees of freedom equal to the number of additional parameters in the model (Haining 1990).

i. Akaike Information Criterion (AIC)

The AIC is measure of fit that is used to assess models. This measure uses the log likelihood with penalizing term that is associated with a number of variables.

Algebraically,

$$AIC = -2\ln L + 2p \quad \text{.....(38)}$$

p is the number of unknown parameters included in the model and lnL is the log likelihood. Smaller values indicate the best model that has to be selected.

ii. Bayesian Information Criterion (BIC)

BIC is another model selection method, which uses penalty term associated with the number of parameters, p and the sample size, n. BIC is also known as the Schwarz Information Criterion

$$BIC = -2\ln L + p\ln n \quad \text{.....(39)},$$

all notations are defined as in (33).

Interpretation: A smaller value shows model with better fit.

Table 4.1 Global Moran's I

Assumption	Observed	Expected	St.dev	Z-value	P-value
Normality	0.51115	-0.0556	0.15552	3.6230	0.0001
Randomization	0.44493	-0.0556	0.147234	3.3992	0.0006

Table 4.2 Global Geary's C

Assumption	Observed	Expected	St.dev	Z-value	P-value
Normality	0.07	1.00	0.0314	3.96	0.00004
Randomization	0.26	1.00	0.2654	-3.021	0.00060

The test results of *Moran's I and Geary's C* describe; the observed (computed) values, expected values, standard deviation, **Z-score**, and the associated p-value. The low p-values ($p < 0.01$) under both assumptions suggest powerful positive spatial autocorrelation. A two-sided p-value is reported, which is the probability that the observed coefficient lies farther away from $|Z|$ on either side of the coefficients; large positive value of **Z** (Morans coefficient) and the large negative **Z** (Geary's C). The coefficient of spatial autocorrelation for both test statistics was significant implying the spatial randomness probability of TB cases is unlikely (<1%).

The Moran scatter plot (Anselin, 1998) is graphical visualization of spatial clustering by displaying the selected (dependent) variable on X-axis and the weighted average of its neighbor areas (lagged values) on Y-axis. The slope of this scatter plot represents the coefficient of Moran's Index and the dots in first and third quadrants indicate positive

spatial autocorrelation (clustered observations), those in second and fourth quadrants show negative spatial autocorrelation (outliers). In Figure 4.2, the scatter plots point out strong positive spatial clustering and therefore nearby woredas have great chance of having related values than distant areas (locations).

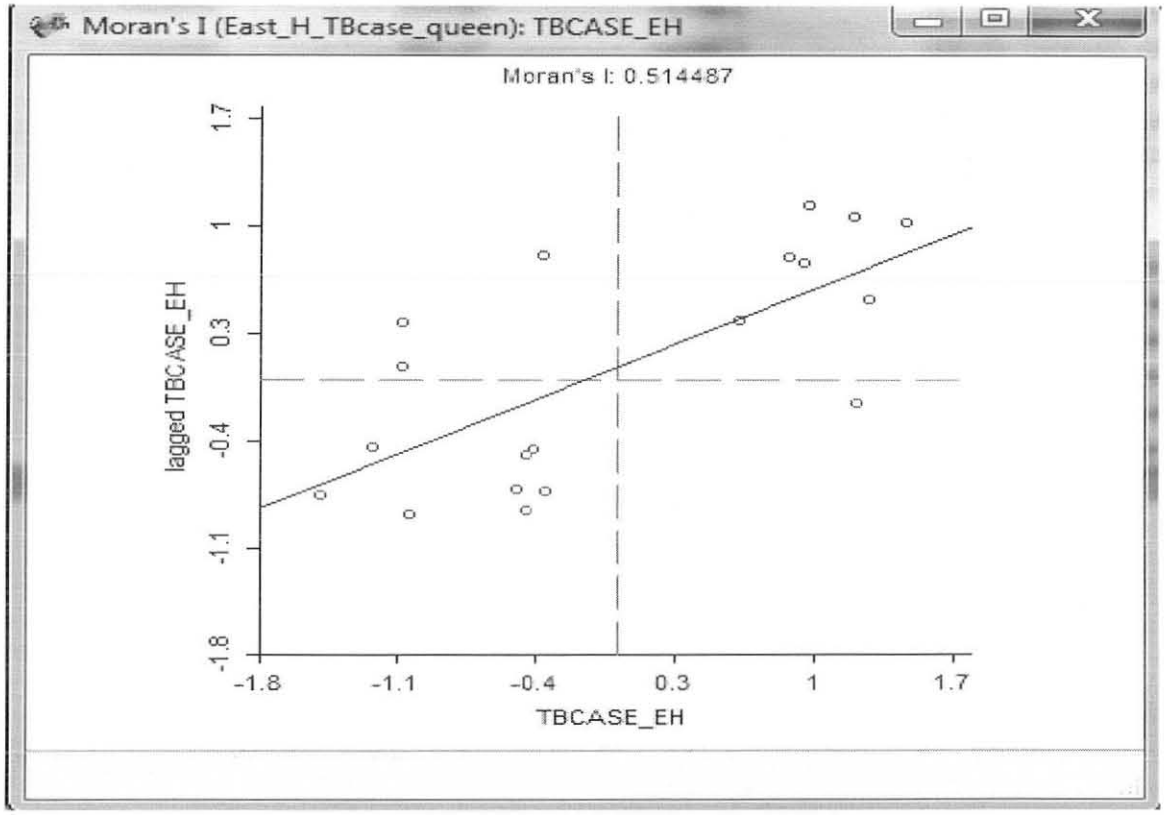


Figure 4.2: SNAPSHOT OF GLOBAL MORAN'S SCATTER PLOT

This scatter plot of TB case versus its lagged values clearly depicts correlation of TB case in neighboring woredas, and this fact strongly supports the results of *Moran's I and Geary's C* statistics. The significance of Moran's scatter plot is commonly tested by taking different large permutation numbers with their respective significance values (see, Figure 1of Appendix).

In general, all the test evidences including; map description (Figure 4.1), statistical test results via global measures, and the scatter plots proved the existence of strong positive spatial dependence as expected. But, this identification of spatial dependence is not the end of the story; we need to explore the spatial pattern of each location for assessing specific risk areas and also for seeking the factors that are contributing for unequal distribution of the cases. For exploring this local spatial pattern *Moran's I* for local spatial statistics and Local *Ord and Getis Gi** (Anselin, 1995) were used.

4.4.2. LOCAL MEASURES FOR SPATIAL AUTOCORRELATION

Spatial cluster of TB cases that was detected by global measures is meaningless unless we identify individual spatial pattern for each woreda, i.e. the test statistics that was rejected with global measures may (may not) be rejected in local case.

Local Moran's I, and *Ord and Getis Gi** test statistics are among the most commonly used spatial test statistics for detecting spatial pattern of areal (lattice) data. The detail description and computation procedure of these test statistics is given in Chapter three (Section 3.3.4). Both local test statistics are used to test against the claim of **No** local spatial autocorrelation for ach woreda in relation to its neighbors. The standardized values of each test statistics are compared to the Z-score and significance of test result follows asymptotic property of normal distribution. The results of Local Moran's index are given in Table 4.3.

Table 4.3: ***Local Moran's I***

S.no	Woredas	*Ii* (LISA)	p-values
1	Chinaksan	0.9561	0.0013
2	Jarso	0.8956	0.0014
3	Kombolcha	0.8802	0.0019
4	Gursum	0.7018	0.0012
5	Meta	0.3624	0.0130
6	Haromaya	0.6726	0.0031
7	Goro gutu	0.5401	0.0025
8	Deder	0.2617	0.0094
9	Kersa	-0.4115	0.1941*
10	Babile	-0.3019	0.1526*
11	Kurfa chele	-0.1805	0.4669*
12	Bedeno	0.2444	0.0062
13	Fedis	0.2459	0.0213
14	Melka belo	0.1874	0.0010
15	Meyu	0.9218	0.00021
16	Kumbi	0.75489	0.00014
17	Girawa	-0.0812	0.5416*
18	Gola oda	0.3957	0.0014
19	Hareri	0.7161	0.0009

Note: The observed values of outliers were negative and insignificant at 5% level.

The test result (Table 4.3) shows significant clustering of TB cases in 15 woredas and outliers in four woredas. This indicates that the null hypothesis of no spatial clustering was not significant for Kurfa-chele, Babile, Girawa, and Kersa at 5 % level of significance. Kurfa-chele has high TB load and was surrounded by small TB cases of neighbors while Babile, Girawa and Kersa were woredas with small TB cases neighbored with woredas of large TB load. The clustering status and the respective geographical location of woredas are given by significance map (Figure 4.3).

Cluster (significance map) is the visual description, and it is simple way of identifying the pattern of spatial distribution than other measures of spatial autocorrelation.

Figure 4.3, presents the map of TB counts of the study area, where each woreda is colored according to the category in to which its corresponding attributes fall. Here, four different colors; red, green, pink and yellow are used. The first two, red and green represent high and low risk areas respectively. As we can see from cluster map, seven woredas which are bordered with Dire-Dawa administration city (North) and Somali Regional State (south) were high areas, account for about 60% of total TB cases.

The test results for Ord and Getis Gi* is presented in Table 1 of appendix, showing clustering of neighboring values; positive Gi* implies clustering of high values, and negative values of Gi* indicate connectedness of small cases (cold spots).

Generally, LISA cluster map has been able to highlight two extreme areas; the high risk area labeled in red and the low risk area labeled in green color. Areas shaded pink and yellow represent outliers (negative spatial autocorrelation).

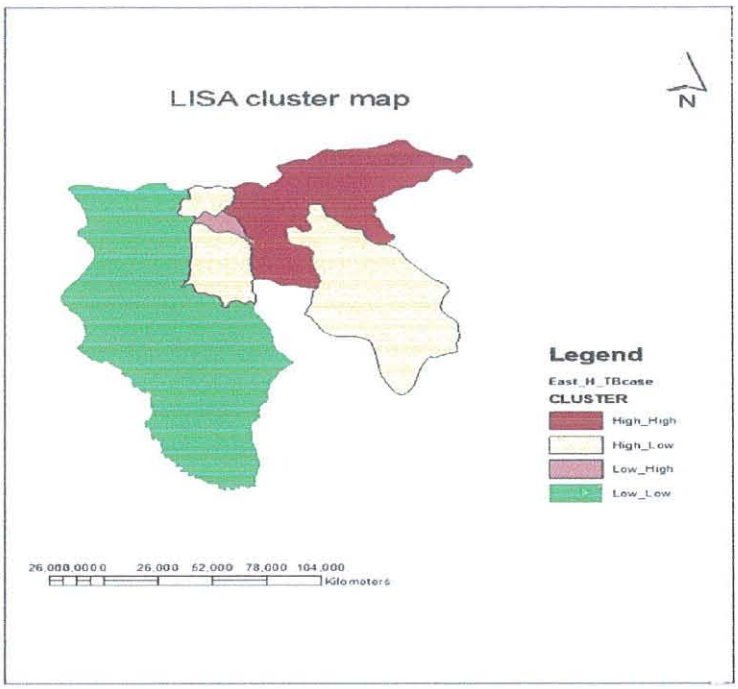


Figure 4.3 LISA CLUSTER MAP

4.5. DIAGONSTIC FOR SPATIAL DEPENDENCE

Tests for spatial autocorrelation, global and local measures supported by cluster map showed the existence of strong spatial autocorrelation for TB load of study area. Once spatial dependence is detected it is not recommended to apply linear regression which assumes error terms have a zero mean and are independently and identical distributed. Spatial autoregressive model is the class of regression model that incorporates spatial dependence as an additional explanatory variable of the model. There are two most common ways to incorporate spatial effects; as lag of dependent variable (spatial lag) and as component of error term (spatial error term). Spatial lag model is considered when the

dependent variable at location i is affected by the independent variables in both i and j, while spatial error model is used in cases where the error term across the spatial units are correlated.

For selecting the appropriate spatial model; first we fit a standard linear regression i.e. ordinary least squares (OLS) model was fitted to determine the linear relationship between TB incidence and explanatory variables, and then we use this model as reference to compare alternative spatial models. Spatial model comparison is based on test statistics called Lagrange Multiplier (LM). The model diagnosis of OLS reports six test statistics; Moran's coefficient for residual, and components of LM test statistics.

Table 4.4: Diagnostic for spatial dependence (using row –standardized weights)

Test	MI/DF	VALUE	PROB
Moran's I	0.43054	8.45049	0.000000
Lagrange Multiplier (lag)	1	6.830074	0.000068
Robust LM (lag)	1	5.232395	0.002221
Lagrange Multiplier (error)	1	1.888823	0.169398*
Robust LM (error)	1	0.290687	0.589826*
LM(SARMA)	2	7.120658	0.028429

***Insignificant at 5% level**

From test result we identify significance of the first three test statistics; Moran's, LM for lag and its robust form.

The value of Moran's I is found to be 8.45 suggesting the existence of strong positive spatial autocorrelation. The first two results of LM test; LM-Lag and Robust LM-Lag pertain to the spatial lag model as the alternative, while the next two; LM-Error and Robust LM-Error refer to the priority of spatial error model. The last LM-SARMA

(higher order alternative) is not useful in practice, but the indication of this test is that, it will tend to be significant when either the error or the lag model is the proper alternative.

The LM test was performed to compare the SLM and SEM alternatives. The test value of SLM was 6.84 ($p<0.001$), implying that the SEM alternative did not account for the spatial dependence as effectively as the SLM model. The coefficient value for LM-SARMA is significant at 5%, suggesting that spatial lag model is a better fit.

After fitting spatial lag model, the assumption of uncorrelated error term is checked by calculating Moran's spatial autocorrelation for lag residuals. As it's visualized in Figure 4.4, the slope (Moran's I for residual of SLM) is near to zero, implying the complete removal of the spatial dependence in spatial lag model.

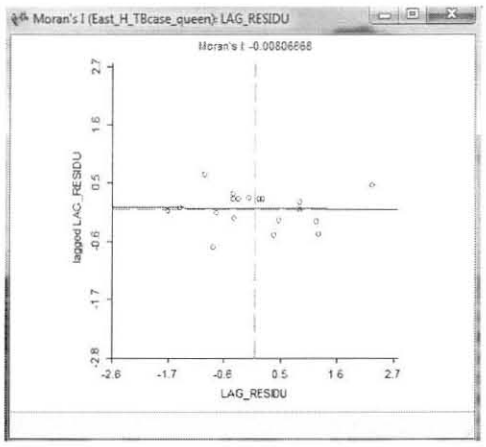


Figure 4.4 snapshots of Mors's scatter plots for error residual

In addition to the LM test statistics there are three common model selection criteria (discussed in methodology, section 3.5) for comparing standard regression and alternative spatial autoregressive models; AIC, BIC and value of log likelihood. The output of model diagnosis from regression models is presented in Table 4.5

Table 4.5: MODEL DIAGNOSTICS SUMMARY

Model	Log Likelihood	AIC	Schwarz Criterion
OLS	-109.248	226.49	230.27
SLM	-103.582	217.16	221.887
SEM	-104.638	219.25	223.037

Table 4.5 shows summary results of model diagnosis for ordinary least square method (OLS), spatial lag model (SLM) and spatial error model (SEM). The model with large likelihood and relatively small values of AIC, and BIC fits better the data. Accordingly, SLM has relatively large log likelihood and small AIC and BIC values, implying the superiority of the spatial lag model to describe the spatial dependence of TB cases on local risk factors (covariates).

4.6. SPATIAL LAG MODEL (SLM)

The spatial lag model is one of the spatial autoregressive models which incorporate the spatial dependence as additional lag dependent variable controlled by the exogenous matrix of spatial weights. Mathematically, it is given by:

$$y = \rho Wy + X\beta + \varepsilon \quad \text{.....(42)}$$

Where y is an (N × 1) vector of observations on a dependent variable taken at each of N locations, X is an (N × k) matrix of exogenous variables, β is a (k × 1) vector of parameters, ε is an (N × 1) vector of disturbances and ρ is a spatial autoregressive parameter for spatial lagged dependent variable.

The spatial effect can be filtered out and equation (42) can be rewritten as:-

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1} \varepsilon \quad \text{.....(43)}$$

The result of model parameters for spatial lag model, processed by Geoda is given in Table 4.6. The output includes coefficients of model parameters, standard error, Z-values, and significance level.

Table 4.6: MLE results for spatial lag model

Variable	Coefficient	Standard error	Z-value	P-value
Constant	1.18	0.358	2.315	0.00215
W_ TB case	0.6965	0.242	3.9605	0.000045
DENS	0.3590	0.04441	8.422	0.00000021
HIV-PREV	0.169	0.0583	2.9603	0.000352

As we can see all model parameters except number of health centers are significant (5%). The result confirms high significance of neighboring TB incidence, population density, HIV prevalence on the spatial distribution of the TB cases.

Therefore, the spatial distribution of TB incidence in East Hararge can be modeled as:

$$Y = 1.18 + 0.6965X_1 + 0.359X_2 + 0.169X_3$$

Where Y is TB incidence of woredas,

X_1_ lagged TB incidence

X_2_ population density

X_3_ HIV prevalence

The spatial model of East Hararge TB distribution suggests strong positive correlation between TB case and proximity, population density, and HIV prevalence.

The contribution of each explanatory variable on TB load can be seen by controlling the effects of others. In this study bivariate spatial autocorrelation is used for detecting this relationship. The bivariate LISA is a straightforward extension of the LISA functionality to two different variables, one for the location and another for the average of its neighbors. Figure 4.5 describes bivariate spatial autocorrelation of; TB case and lagged HIV prevalence, TB case and lagged population density.

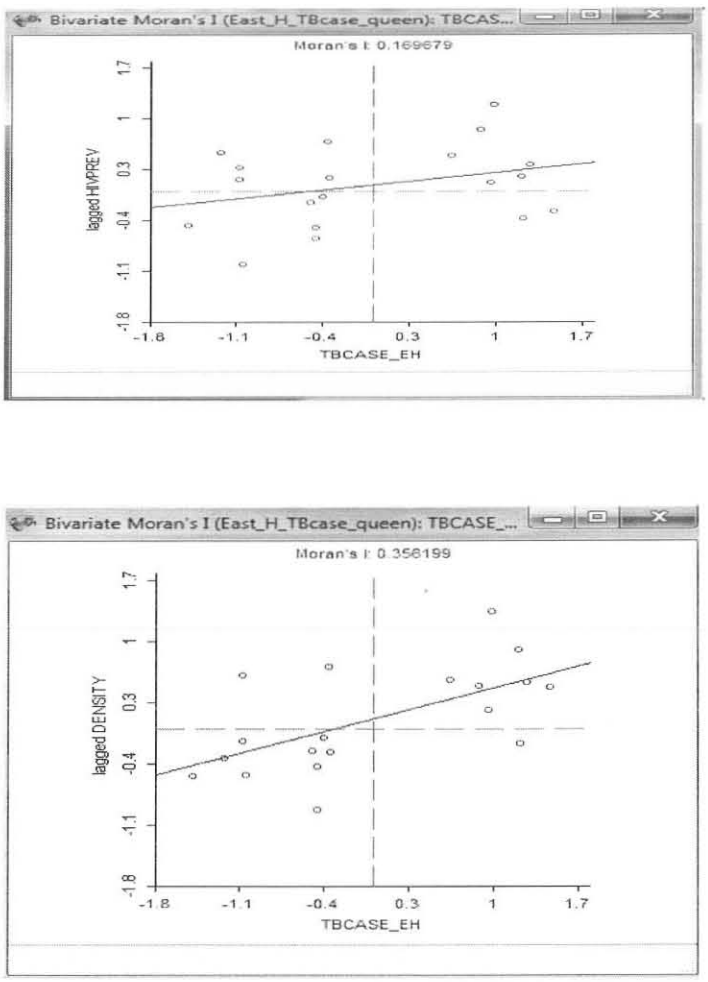


Figure 4.5 Bivariate spatial autocorrelation

Figure 4.5 shows the spatial association between TB cases and local risk factors. As we can see, the spatial association between TB cases and population density is large while TB case has small direct correlation with HIV prevalence.

Regression diagnostics for spatial lag model

The Breusch-Pagan and Jacques-Bera were used to test the null hypothesis against the constant variance and normality of error terms respectively. Summary output of regression diagnostic for spatial lag model is given in Table 4.7

Table 4.7 **Result of regression diagnostic for spatial lag model**

Test statistics	DF	Value	prob
Bruesh-pagan	3	3.52	0.318
Jarque-Bera	2	1.425	0.4951
Likelihood ratio test	1	0.1999	0.6545

Result of regression diagnostic showed that the problem of heteroskesdasticity and non normality of error term completed removed with spatial lag model; we can use Maximum Likelihood estimate for parameter estimation. LRT revealed that, spatial effect has been removed with spatial lag dependence, implying that spatial lag model is appropriate model to describe the linear association of TB incidence with local risk factors.

4.7. BAYESIAN CONDITIONAL AUTOREGRESSIVE MODEL (CAR)

The primary purpose of CAR model is to provide a modeling mechanism to account for residual autocorrelation, spatial trend that is not explained by spatial patterns in covariate value (Waller and Gotway, 2004).

Given the structure of the CAR specification, it is necessary to know the neighbors of each region, and by including covariates in our model we aim to assess and remove the effect of potential confounders or risk factors.

The assessment of the importance of a covariate is indicated by the estimated value of its coefficient and its associated probability interval. If, for example, the 95% credible interval does not contain the value 0, it means that the coefficient is significant and, if greater than zero, it will indicate a positive relationship between the risk and vice versa.

4.7.1. THE CAR MODEL AND PRIOR SPECIFICATION

The CAR model we used for describing spatial distribution of TB cases has three components: the likelihood for count of TB data, the latent process model for log relative disease risk and the prior specification for model coefficients, and spatial error term.

Gibbs sampling has been used for simulating model parameters of Poisson- CAR based on; observed count, expected count and covariate of TB cases with adjacency matrix defining neighborhood structure of woredas.

CAR MODEL FOR TB DATA

MODEL

```
model {
# Likelihood
for (i in 1 : N) {
O[i] ~ dpois (mu[i])
log (mu[i]) <- log(E[i]) + alpha[0] + alpha[1]X1[i] +beta*X2[i]+gamma*X3[i]+ b[i]
# Area-specific relative risk (for maps)
RR[i] <- exp (alpha[0] + alpha[1] * X1[i]+beta*X2[i]+gamma*X3[i] + b[i])
}
# CAR prior distribution for random effects:
b[1: N] ~ car.normal ( adj[ ], weights [], tau)
for (k in 1:sumNumNeigh){
weights[k]<-1
}
```

4.7.2. POSTERIOR ANALYSIS

The posterior estimation of the model parameters was simulated by Gibbs sampling, where the convergence assumption achieved after iteration (20,000), and the calculated values are given in Table 4.7.

The calculated deviance information criterion (DIC) for Poisson-CAR model was 151.7, and the 95% credible intervals for covariates does not include for all model parameters, implying significance of all local risk factors in determining spatial distribution of TB cases.

Table 4.7: POSTERIOR STATISTICS FOR CAR MODEL

Parameter	mean	Sd	MC-error	2.5%	97.5%
$alpha_0$	-39.57	1.734	0.145	-42.64	-37.42
$alpha_1$	0.426	0.058	0.004	0.3262	0.504
beta	-0.064	0.014	0.0012	-0.088	-0.0012
gamma	0.035	0.006	5.11E-4	0.0181	0.051

The coefficients of CAR models justified the significant contribution of local risk factors; population density, number of health centers and HIV prevalence, and the result of CAR models is nearly the same as spatial lag model (Table 4.6), implying appropriateness of both CAR and SAR in modeling spatial distribution of TB cases.

4.8. Summary

Spatial analysis was carried out in two steps to identify and quantify the spatial distribution of TB cases in East Hararge Zone including Harari Regional State which is engulfed by East Hararge Zone. Firstly, autocorrelation analysis was carried to examine the spatial pattern of TB cases among neighboring woredas of study area, and secondly, two different autoregressive models (spatial autoregressive model and conditional autoregressive model) were fitted in order to describe the association of TB case with local risk factors.

The test results of spatial autocorrelations revealed strong association of nearby values; significant clustering of 15 neighboring woredas, of which seven were founded to be high risk areas. The cluster map (Figure 4.2) identified two extreme areas; clustering of high values on north and eastern part, and condensation of similar small values on western region of the study area. Furthermore, both spatial autoregressive (spatial lag) and Bayesian conditional autoregressive (CAR) models reflected spatial relationship of TB incidence (load) with local risk factors.

4.8.1. DISCUSSION OF THE RESULT

This study was intended to identify whether the prevalence of TB is independent of location or geographical factors. The study was conducted based on all forms of TB data (2004 E.C). The result of study revealed the association between TB cases and local risk factors: proximity to affected groups, population density, HIV prevalence, and number of health centers.

A study conducted in North Shoa, Ethiopia revealed the strong association of TB load to the number of health center, population density and HIV prevalence (Habte T, 2011). Similar study in Greater Banjul, Gambia (Touray.et.al, 2010) showed population density as major determinants of spatial clustering of TB cases. Another study by Omoloke (2012), revealed the effects of socio-economic factors, especially Health facilities. For example; the difference between Nigeria of Africa and UK of Europe in preventing and controlling the prevalence of TB and HIV cases. The study conducted in Malawi (Nyirenda, 2005) suggest the strong spatial effect of local risk factors like, poverty, HIV infection, household contact, and overcrowding on spatial distribution of TB cases.

This study employed two different spatial models for describing the spatial pattern of TB cases in study area. Models of the study, spatial lag and conditional autoregressive showed significant spatial dependence of TB cases in neighboring woredas. Venkatesan and Srinivasa (2010) used three different models; Spherical, exponential and Gaussian for modeling the spatial variogram of the TB cases in Chennai districts, India. Accordingly, the results of their study revealed strong association of TB cases in nearby woredas, or woredas with small distance from each other.

Similar studies on spatial distribution of TB cases; Thomas and Richard, 2004, and Munch.et.al, 2003 identified the significant contribution of demographic and socio-economic factors for unequal distribution of TB cases. Gender, sex, crowding and age were identified as significant factors for spatial distribution of TB.

This study showed that TB has a positive spatial autocorrelation with HIV; the effect of HIV prevalence on spatial distribution was statistical significant. On other hand, a study conducted in Switzerland, (Sudre.et.al, 1996) revealed insignificance of HIV prevalence on spatial distribution of the TB

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

This study utilized exploratory spatial analysis to identify and characterize the spatial distribution of TB cases in East Hararge Zone, Oromia Region including Harari Regional state, Ethiopia.

5.1. Conclusion

This study revealed strong positive spatial clustering of TB cases in study area; nearby woredas have similar values of TB cases than distant woredas. Specifically, two extreme areas were identified by local spatial statistics. High-risk areas occurred in seven neighboring woredas geographically located at the upper- north and eastern region of the study area, and low risk (cold spots) on western part.

After detecting significant spatial clustering, the study employed two different spatial models; spatial autoregressive model (SAR) and Bayesian conditional autoregressive model (CAR). Both CAR and SAR analysis indicated significant effects of local risk factors on spatial distribution of Tuberculosis.

The overall study findings suggest that TB prevalence of study area is highly enhanced by proximity to affected areas, and high population density and HIV prevalence and number of health centers.

5.2. Recommendations

Based on the result obtained the study recommends that the concerned bodies (health planners) should give special attention for high risk area where TB prevalence was very large and clustered, and take appropriate intervention.

The study also recommends for other researchers to conduct spatial analysis of TB with small scale aggregation like, kebele or house hold level.

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. Anselin, L. (1992). "SpaceStat Tutorial: A Workbook for Using Space Stat in the analysis of Spatial Data." Typescript. University of Illinois at Urbana-Champaign, pp. 8-67.
3. Anselin, L. (1995). Local indicators of spatial association - LISA. *Geographical Analysis*, **27**: 93-105.
4. Anselin, L. (1998). ''Interactive Techniques and Exploratory Spatial Data Analysis.'' In Geographical Information Systems: Principles, Techniques, Management and Applications, pp. 251–64
5. ANSELIN, L. 1998. Spatial econometrics. Available online at: www.csiss.org/learning_resources/content/papers/baltchap.pdf
6. Anselin, Luc (1988) Spatial Econometrics: Methods and Models, Dordrecht. *The Netherlands: Kluwer Academic Publishers.*
7. Assel Terlikbayeva, Sabrina Hermosilla, Sandro Galea, Neil Schluger, Saltanat Yegeubayeva, Tleukhan Abildayev, Talgat Muminov (2012): Tuberculosis in Kazakhstan: analysis of risk determinants in national surveillance data. *BMC Infectious Diseases*, **12**:262
8. Besag J (1974) spatial interaction and the statistical analysis of latice systems. *J. Roy. Statist. Soc. Ser. B* **36** , 192-236.
9. Burridge, P. (1980). Testing for a common factor in a spatial autoregressive model. *Environment and Planning* , **A 13,**795-800.
10. Carla Nunes (2007): Tuberculosis incidence in Portugal: spatiotemporal clustering. *International Journal of Health Geographics*
11. Clayton D, Kaldor J. (1987) *Empirical Bayes estimates of age-standardized relative risks for use in disease mapping Biometrics,* **43**, 671–681.
12. Cliff, A. and Ord, J. (1981): Spatial processes, Modeling and application. *Pion, London.*

13. Cliff, A.D. and J.K. Ord (1972). Testing for spatial autocorrelation among regression residuals. *Geographical Analysis,* **4**, 267-84.
14. Cressie, N. (1993): Statistics for spatial data. *Wiley, New York*
15. Dirk Pfeiffer (2008) spatial analysis in epidemiology. *Oxford university press,GB.*
16. Dr. Thomas C. Schafer and Dr. Richard Lisichenko (2004): Spatial Variations of Tuberculosis In The State Of Kansas, USA.
17. *Federal Ministry of Health (2008): Implementation Guidelines for TB/HIV Collaborative Activities in Ethiopia.*
18. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: the Analysis of Spatially Varying Relationships. *John Wiley and Sons*
19. Friedrich, G. (1998): Survey on cluster tests for spatial area data. *Computation Statistics and Data Analysis,* **31**, 39-58.
20. Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistics* **5**, 11545.
21. Gesler W (1986). The uses of spatial analysis in medical geography. *a review. Soc Sci Med*, **23,** 963-73.
22. Getis, Arthur (1990). Screening for spatial dependence in regression analysis. Papers, *Regional Science Association*: **69**, 69-8
23. Habte T (2011). Statistical analysis of spatial distribution of TB in NorthShoa, Ethiopia. *Department of Statistics, unpublished MSc thesis.*
24. Haining, R.P. (1990). Spatial data analysis in the social and environmental sciences. *Cambridge University Press, Cambridge,UK.*
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 Master thesis)*
26. Jamshid. Y.Ch. and A. Kazemnegad (2010). Spatial Distribution of Tuberculosis in Mazandaran Province–Iran: Spatiotemporal Modeling
27. Kline, S.E., L.L. Hedemark, S.F. Davies. 1995. Outbreak of Tuberculosis among Regular Patrons of a Neighborhood Bar. *The New England Journal of Medicine,* **333(4)** 222-25.

28. Kulldorff, M. and Nagarwalla, N. (1995): Spatial Disease Clusters, Detection and Inference. *Statistics in Medicine,* **14,** 799-810.

29. Longley P, Goodchild M, Maguire D, Rhind, M (1999): Geographical Information Systems: Principles, Techniques, Applications and Management. *Wiley, New York.*

30. Manfred M., Fischer, Jinferg Wang: modeling area data, in spatial data analysis. *Springer briefs in regional science*, pages 31-44. *Springer Berlin Heidelberg*, 2011.

31. Marshall (1991). Mapping Disease and Mortality rates using Empirical Bayes Estimators. *Journal of the Royal Statistical Society, Series C, Applied Statistics.*

32. 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 Geographic's*, **24**, 3-25.

33. Matthews SA (1990). Epidemiology using a GIS: the need for caution *Compute Environ Urban Syst*, **14:**213-21.

34. Mayer JD (1983). The role of spatial analysis and geographic data in the detection of disease causation. *Soc Sci Med*, **17**, 1213-21

35. Ming-Jung Ho (2004): Socio cultural aspects of tuberculosis: a literature review and a case study of immigrant tuberculosis, *Social Science & Medicine* **59**,753–762.

36. Mollie A (1999). Bayesian and Empirical Bayes Approaches to Disease Mapping. In: Lawson A, Editor. Disease Mapping and Risk Assessment for Public Health, *John Willey and Sons Ltd.*

37. M. Wall (2004): a close look at the spatial structure implied by the CAR and SAR models. *Journal of statistical planning and inference*, **121(2):**311-324.

38. Nguyen Thi Vananh (2012): Molecular epidemiology of Tuberculosis in Vietnam (2003-2009)

39. ORD, J.K. 1975. Estimation methods for models of spatial interaction. *J. Am. Stat. Assoc.* **70**:120 –126.

40. Ord, J. and Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, **27** (4): 286-306.

41. P. Sudre, B. Hirschel, L. Toscani, B. Ledergerber, H.L. Rieder, and the Swiss HIV Cohort Study. Risk factors for tuberculosis among HIV-infected patients in Switzerland. *ERS Journals Ltd* ,1996.

42. P. Venkatesan and R. Srinivasan(2010): Modeling the spatial variogram of tuberculosis for Chennai ward in India. *TuberculosisResearch Centre, ICMR, Chennai* – **600**, 031, India

43. Semeeh A. Omoleke(2012): An analysis of tuberculosis in developing and developed world: Nigeria and UK as a case study. *Journal of Public Health and Epidemiology*, **4(6),** pp. 150-155.

44. Thomas Nyirenda (2005). Epidemiology of Tuberculosis in Malawi

45. Tobler W (1979): Cellular geography. In: Gale S, Olsson G, eds.Philosophy in geography. Dordrecht, *The Netherlands: Reidel*, 379-86.

46. 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.

47. Venkatesan P and Srinivasan R. (2008): Applied Bayesian statistical Analysis. *Proceeding of NSABSA* ,pp:51-56

48. Waller, L. A., & Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data: *John Wiley & Sons, Inc.*

49. WHO (2009) Global Tuberculosis Control: A short update to the 2009 report, *WHO Technical Report.*

50. WHO (2007): Treatment of tuberculosis. *Guidelines for National Programmes. Geneva*

51. William Brennan Arden (2008). Spatial and temporal-spatial clusters of tuberculosis in ceara state, brazil, using gis and the scan statistic.

52. Z. Munch,* S. W. P. Van Lill,* C. N. Booysen,* H. L. Zietsman,† D. A. Enarson,‡ N. Beyers*,Tuberculosis transmission patterns in a high-incidence area: *a spatial analysis in South Africa.* (2003)

APPENDIX

Figure1: PERMUTATION TEST RESULT FOR GLOBAL MORAN'S I

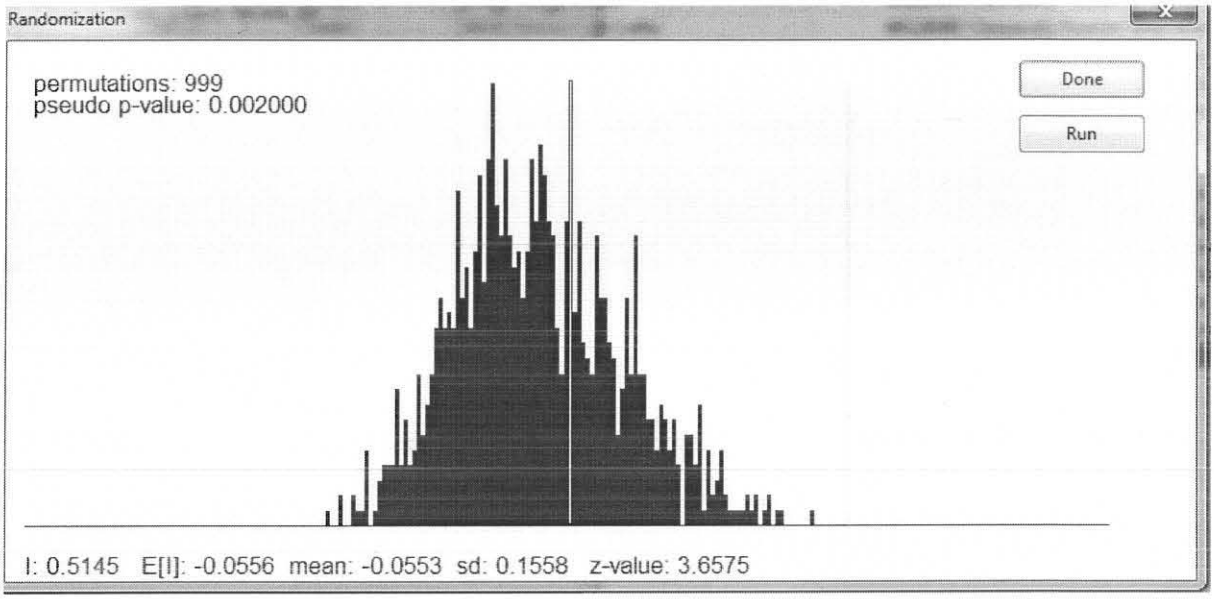


Figure 2. BIVARIATE MORAN'S SCATTER PLOT

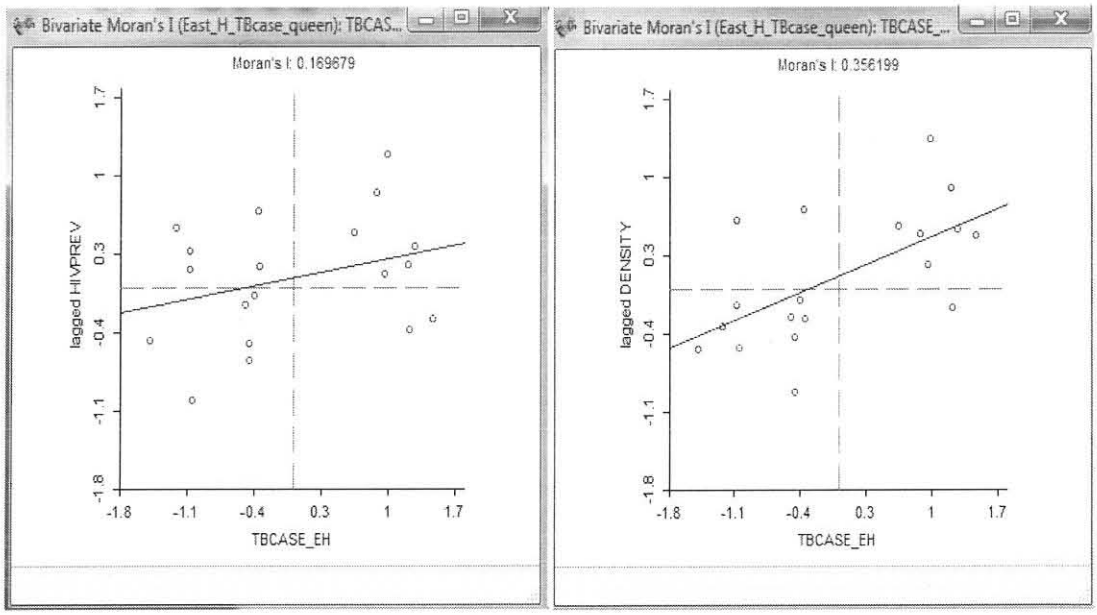


Table2. SOCIO-DEMOGRAPHIC VARIABLES AND TB LOAD OF WOREDAS

S.no	W-name	Area(KM^2)	Pop-size	Pop_den	Expected (pop-size)	TB_case
1	Babile	5424.73	179694	33.125	223	150
2	Bedeno	1041.25	209551	201.25	267	140
3	chinaksan	1494.96	246922	165.17	307	352
4	Deder	512.66	189750	266.41	236	151
5	Fedis	1026.13	228047	222.24	284	260
6	Girawa	1443.26	66750	46.25	83	71
7	Gola-oda	1861.45	280229	325.3	348	140
8	Goro-gutu	491.23	96290	196.02	119	54
9	Gursum	793.18	216549	273.014	269	287
10	Harari	371.28	183505	494.25	228	295
11	Haromaya	533.99	217467	407.25	270	331
12	Jarso	518.85	131840	254.102	164	323
13	Kersa	449.54	55222	85.0181	69	71
14	Kombolcha	468.69	123242	262.95	153	298
15	Kumbi	2032.54	18647	18.06	23	26
16	Kurfa-chele	268.11	179287	383.003	223	325
17	Mayu	4955.42	59663	12.04	74	75
18	Melka-belo	1404.91	143342	102.03	113	145
19	Meta	655.73	91505	139.548	178	135
	Total	25747.91	2,917,502			3629

Table 3. Contiguity weigh matrix (Queen's-method) for east Hararge woredas

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
4	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1
6	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1
7	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	0	0
9	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
11	0	0	0	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0
12	0	0	0	0	1	0	0	0	1	0	1	0	0	1	1	0	1	1	0
13	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	1	0	1
14	0	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0
15	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
17	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0
19	0	1	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0

Note: the serial number given for weight matrix is defined according to the respective number given in Table 1 of appendix. (Ord and Getis, Gi* statistics).