TITLE: SPATIOTEMPORAL CLUSTERING AND MODELLING OF TUBERCULOSIS IN KENYA.

A Project report

Submitted by

Margaret Lilian Ndubi

P.O.Box 63037, 00100

Nairobi, Kenya

Telephone number-+254722293149

e-mail-ndubimeg@yahoo.com

In partial fulfilment for the award of the Degree

Of

Masters of Science in Medical Statistics

At

University of Nairobi Institute of Tropical and Infectious Diseases

(UNITID)

Signature ---------- Date----------

November 2011

Declaration

I hereby declare that this research project is my original work and has not been presented for Academic award in any other university.

Name: Margaret Lilian Ndubi

Signature:----------

Date: 25th November 2011

This research project has been submitted for the award of the degree of Masters of Science in Medical statistics with my approval as the University supervisor for Margaret Lilian Ndubi.

Name: Dr Thomas N.O. Achia

Title: Dr

Institution of affiliation: University of Nairobi

Address: P O Box 30197-00100 Nairobi

Signature: 

Date: 25th November 2011

Dedications

It is my pleasure to dedicate this project to my parents, my siblings and friends .I also want to dedicate it to the public especially those interested in spatial statistics with the hope that it will improve their knowledge on spatial and spatiotemporal techniques.

Acknowledgement

I wish to acknowledge The Almighty God for giving me good health and provision throughout this period. My supervisor Dr Thomas Achia for the direction and assistance he offered during the project. The division of leprosy, Tuberculosis and lung diseases for their support and allowing TB data to be used in this project. I also want to acknowledge my fellow class mates and colleagues at work for their tireless continued support and encouragement.

Abbreviations

DLTLD	Division of Leprosy, Tuberculosis and Lung Diseases.
DOTS	Directly Observed Therapy, Short course
DTLC	District Tuberculosis and Leprosy Coordinator
EB	Empirical Bayes.
GIS	Geographic Information System
HIV	Human Immunodeficiency Virus.
KIHBS	Kenya Integrated Household Budget Survey
LISA	Local Indicator of Spatial Autocorrelation.
MDGs	Millennium Development Goals
MOPHS	Ministry of Public Health and Sanitation
PTLC	Provincial Tuberculosis and Leprosy Coordinator.
TB	Tuberculosis
W.H.O	World Health Organization.
SMR	Standardized Morbidity Ratio
CAR	Conditional Autoregressive model
MCMC	Markov Chain Monte Carlo.

Table of contents

List of figures

List of Tables

Abstract

Tuberculosis is an infectious disease caused by the bacteria bacillus mycobacterium tuberculosis and is spread through the air by persons suffering from it, especially those that are sputum smear positive. Kenya is among the 22 countries burdened by TB in the world and reports about 100,000 cases of this disease every year. The aim of this study was to investigate the spatiotemporal clustering of tuberculosis and the factors associated with it in Kenya using Bayesian approach to map and measure spatial variation of TB risks.

The TB data used in this study was obtained from DLTLD from 2002 to 2009 while the data on covariates were obtained from different sources. Combinations of approaches were then used in the analysis. First exploratory spatial data analysis in terms of maps were generated using Bayesian approach then clustering of disease in different districts computed using Moran's I and space time scan statistics. Finally Bayesian approach with CAR specification was computed to identify spatial variation of TB and covariates such as prevalence of HIV, distance greater than five kilometres to the nearest health facility, proportion of the poor in each district, illiteracy rate in each district and urbanization among others.

Results have shown us the spatial distribution of TB and areas that have consistently reported high relative risks over the years such as Nairobi, Kisumu, Bondo and Rachuonyo among others. The effects of the covariates have also been demonstrated to be positive with credible interval that does not contain zero. For example, in HIV the posterior density of β had a posterior median that was greater than zero and a 95% credible interval of 0.200 to 3.157 which was considered statistically significant.

In conclusion, this study revealed the spatial distribution of TB in the country over the years from 2002 to 2009. It also revealed the spatial patterns of TB in different districts in comparison to their nearest neighbouring district, emerging districts with increased TB risk and those districts that have consistently recorded high relative risk of TB. This information can be used by the DLTLD for planning purposes, allocation of resources' and even dissemination of TB information. It can also be used to strengthen other strategies such as poverty eradication in Kenya.

1.0 INTRODUCTION

Tuberculosis is an infectious disease caused by the bacteria bacillus mycobacterium tuberculosis and is spread through the air by persons suffering from it, especially those that are sputum smear positive. The disease is spread through coughing, spitting, sneezing, laughing and talking among others. In Kenya, about 80 percent of the population have latent TB which manifests itself when the immune system is suppressed. In addition, Kenya is among the 22 countries in the world burdened by TB and is ranked 13th among those countries. In Africa alone, it is ranked 5th most affected by the disease, WHO (2009). In the country, TB is a major cause of morbidity and mortality with about 100,000 cases reported every year, DLTLD, MOPHS (2009).

According to DLTLD, MOPHS (2009), TB has been on a steady rise since early 1990's and these increase posses a major threat to the countries health and economy. In addition, case notification has been on the increase from 53 per 100,000 populations in 1990 to 326 per 100,000 populations in 2009. The identified age group with the highest TB case notification rate for both sexes remained 25-34 years in 2009 as indicated in previous DLTLD annual reports.

Report on Global burden of disease, WHO(1990, c.f. Tiwari et al 2006), showed that tuberculosis was ranked 7th among the causes of morbidity in the world and it is not surprising that reducing TB incidence by half in the year 2015 is one of the set targets for the United Nation Millennium Development Goals (MDG).Stop TB partnership, an initiative of WHO endorsed by world health assembly in 2006, aims at halving the prevalence and death rates as a result of TB in comparison to the 1990 levels and targets at eliminating TB (reducing the global active TB incidence to less than one case per a million population per year) by 2050 , WHO(2008).

WHO (2010) and DLTLD , MOPHS (2009) also showed that Kenya as a country has achieved the WHO target of TB case detection rate of 70 percent with a treatment success rate of 85% among the sputum smear positive pulmonary TB cases. Since 80% cases of TB are being detected, there are still 20% undetected TB cases that continue to transmit the infection. In addition, the report results suggested that as much as the number of new cases of TB seems to be declining, the number of those requiring re-treatment is increasing.

WHO introduced Directly Observed Treatment Short course (DOTS) as a global TB control strategy, but this has not been very fruitful in all setups where after its implementation, TB incidence continued to rise. DOTS as a strategy concentrate more on early TB case detection and successful treatment as opposed to prevention. TB as an airborne disease is known to be associated with a number of factors that contribute to its development and or occurrence. As much as early detection and treatment is crucial in the fight against TB, the determinant factors that play a role in its development should be addressed if the disease is to be eliminated. Some of the determinants include-: HIV status, poverty and overcrowding, unemployment, tobacco smoking and indoor air pollution, age, gender, low education status and environmental factors like Distance to the nearest health facility and altitude.

New approaches in GIS such as geostatistical mapping, cluster detection techniques and spatial temporal analysis may contribute significantly in the control of TB through visualization and identification of TB spatial patterns which can be related to the contributory factors and measures to tackle them addressed. This study used Bayesian technique to map Standardized Morbidity Ratio (SMR) of TB in the different districts in Kenya and associated areas with high TB relative risk to the determinant factors. Apart from that, clusters and hotspot areas were detected with the aim of enlightening the Division of Leprosy Tuberculosis and Lung Diseases (DLTLD) so that while planning more resources and emphasis can be directed towards the areas that are most affected. In addition spatial outliers in terms of TB occurrence were also detected i.e. areas with low relative risk of TB surrounded by areas with high relative risk of TB and vice versa.

1.1 Main objective of the study

To investigate the spatiotemporal clustering of tuberculosis in Kenya and the factors associated with it.

1.1.1 The specific objectives of the study are:

a) To review and use appropriate spatial smoothing techniques in order to reduce instability due to the high variability of the tuberculosis data in the different districts in the country.
b) Identify tuberculosis hotspot districts in the country through maps and space time scan statistics.

c) Use spatial statistical models to link the occurrence of tuberculosis to its determinants

1.2 Hypothesis.

Assuming constant risk hypothesis in the spread of TB, there is no significant difference in TB incidence among the different districts in the country.

1.3 Justification.

Since tuberculosis is a major cause of morbidity and mortality in Kenya and in the world at large, a number of efforts have been put in place to reduce the incidence of this disease with a goal of reducing it to less than 1 case per a million populations per year by the year 2050. Achievements have been made so far but unfortunately the reduction is slow as it is projected that about eight of the high burdened countries will not achieve the MDG TB target for 2015, though Kenya has a chance of achieving the set target. In addition, according to WHO (2010), it has been noted that the incidence of TB is falling slowly while the notification rate is increasing in the African region. In view of the above, it was important that spatial patterns and factors associated with tuberculosis be unveiled so that other strategies apart from DOTs can be emphasized to enhance and or improve TB reduction process. Since TB disease patterns and the associated factors can be revealed through spatial and spatial temporal analysis, the study therefore aimed at looking at the spatiotemporal clustering and modelling of TB with more emphasis on the determinant factors that hinder or hamper the reduction of this disease. Some of the factors considered were-: HIV, poverty, Household size, unemployment, indoor air pollution, illiteracy rate, distance to the nearest health facility and altitude.

2.0 LITERATURE REVIEW.

2.1 Spatial smoothing and Mapping

In public health, the goal of mapping is to provide an insight of the geographic occurrence of disease or variation in disease risk and or rates. The maps are also used to identify and understand the cause of diseases, in policy formulation and allocation of resources needed to combat or reduce the spread of the disease identified. It is however not meaningful to have maps of counts because areas with higher population are expected to have increased disease occurrence as compared to areas with lower population, and as a result the population can be accounted for by using rates or SMR. Unfortunately if the population sizes are very different, this may still obscure the spatial patterns and one of the solutions to this problem is spatial smoothing, Waller and Gotway (2004).

Spatial smoothing is the process of reducing the noise in rates or SMR associated with different geographic region. This is accomplished by borrowing information from the neighbouring region so as to produce a better estimate of the rates or relative risk associated with each region and separates the spatial pattern from the noise. It comes in handy in small number disease where disease rates tend to be extremely unreliable because of small numbers upon which it is based.

In a study by Zaman et al (2006), spatial rate smoothing technique was employed in order to reduce noise so that distinct patterns of TB prevalence can be observed in rural Bangladesh. Fang et al (2006) also implemented spatial rate smoothing while mapping annualized Haemorrhagic Fever with Renal Syndrome (HFRS) in China from 1994 to 1998. Kazembe (2007) used Bayesian technique to map malaria risk in northern Malawi. Furthermore, Callaghan et al (2009) applied Bayesian smoothing technique and compared it to classically calculated rates (standard rates) where the result suggested that smoothing gives distinct patterns of disease in a map as compared to rates or proportion.

The empirical Bayesian smoothing approach was also applied by Uthman et al (2009) when analysing the trend and the distribution of TB-HIV deaths in Africa from 1991 to 2006. They used yearly data submitted to WHO on deaths related to TB-HIV and modelled using multilevel Poisson growth approach with more emphasis on trends. Bayesian smoothing

approach was also used in Antananarivo a city in Madagascar by Randremanana et al (2010) to map risks of TB.

2.2 Spatiotemporal models

Spatiotemporal models are approach used when data is collected over space and time. These approaches do not detect the cause of disease but rather provides an insight of the areas with high incidences which can be investigated further for causal relationship. In addition, the method can identify the source of disease occurrence with respect to demographic, time and space. Spatial temporal methods are useful in monitoring of disease status in a community as they may provide information on changing disease pattern over time.

Tiwari et al (2006) investigated on the geo-spatial hot spots for the occurrence of tuberculosis in Almond district, India, using GIS and spatial scan statistics. They identified areas with high incidences of TB. In Portugal, Nunes (2007) identified Tuberculosis clusters using spatiotemporal techniques. He specifically used space time scan statistics to spot most significant clusters in the study area. Onozuka et al (2007) on the other hand, predicted clusters of tuberculosis in Fukuoka area in Japan from 1999 to 2004 using space time scan statistics. They utilized Monte Carlo simulation set at 999 replication and reported clusters of TB with a statistical significance level of $p<0.05$. The same technique of cluster detection was employed by Randremanana et al (2009) in Antananarivo a city in Madagascar to identify TB clusters. In Gambia, 80% of TB cases were found in western part of the country known as Greater Banjul. Due to these high incidences of TB, Touray et al (2010) decided to conduct a study in order to reveal hotspots in this particular area. They used aggregated data to perform spatial scan statistics that revealed significant TB clusters both spatially and over space and time.

2.3 Tuberculosis determinants

TB as a communicable disease is associated with several factors as key determinants or risk factors. The factors can broadly be classified as-: socio-economic, socio-demographic and environmental factors. In addition, HIV /AIDS, diabetes, Cigarette smoking, alcohol and drug abuse are also considered major contributors of increased TB incidence.

Studies have shown that socioeconomic factors that lead to poverty contribute significantly to the development of TB in that the poor live in crowded areas, have poor nutritional status

(malnutrition) due to poor eating habits and or lack of food that lead to weak immune system. They also have limited access to proper medical care (diagnosis and treatment), Waaler et al (2002). Barr et al (2001) demonstrated that TB incidence rose with increasing neighbourhood poverty and overcrowding. In the same vein, Munch et al(2003) investigated the risk factors and areas of TB transmission using GIS and revealed a significant association of TB with unemployment, overcrowding and number of sheebens (crowded neighbourhood bars). In addition, Crampin et al (2004) had similar results since increased TB incidences were observed in those who had contact with TB case, those with fewer possessions, those who shared sleeping dwellings and the ex-drinkers.

Chan-Yeung et al (2005) discovered that poverty, low education attainment and old age were significantly associated with high rates of TB whereas Baker et al (2007) also reported significant association between TB and household crowding. Furthermore, Dye et al (2009) in their study of determinants and trends of TB incidence in 134 countries found out that TB incidence was declining faster in countries with improved sanitation, low child mortality and high human development index. Urbanization on the other hand has lead to mushrooming of slums and overcrowding. These predisposes individual to TB cases thus increased TB incidence, Randremanana et al (2010)

Environmental factors also contribute to the likelihood of acquiring the infection because the concentration of TB bacilli depends on the ventilation of the surroundings and exposure to ultraviolet light because Vitamin D helps with macrophage function, and macrophages help to clear TB bacteria. Thus, overcrowding, congregation in prison settings, poor housing, and inadequate ventilation increases the risk of contact and predisposes individuals to the development of TB. Lienhardt et al (2005) and Hill et al (2006) proved that family history or previous exposure to TB from a household member is a key factor in developing the disease. In addition, overcrowding, history of being in prison, unemployment and the use of illicit drugs as TB risk factors were demonstrated by Coker et al (2006). Distance to the nearest health facility affect health seeking behaviour and access to health care. Thus the further the health facility, the poorer the access by especially those of low socio economic status. Since TB is an airborne disease, the more a person stays with the untreated disease, the more it spreads. Randremanana et al (2010) showed a positive relationship between TB and the distance to the nearest health facility, though the relationship was not significant.

Socio-demographic factors such as age, gender and ethnicity have been associated with TB. Studies conducted by Lienhardt et al (2005) revealed that the risk of TB increased with age, was higher in the male gender and single individuals including those who are widowed or divorced as compared to the female gender and the married individuals. The results are similar to those of Crampin et al (2004) who had identified that being widowed separated or divorced increased the risk of TB. On the other hand, ethnicity as a factor was described by Hill et al (2006) in Gambia.

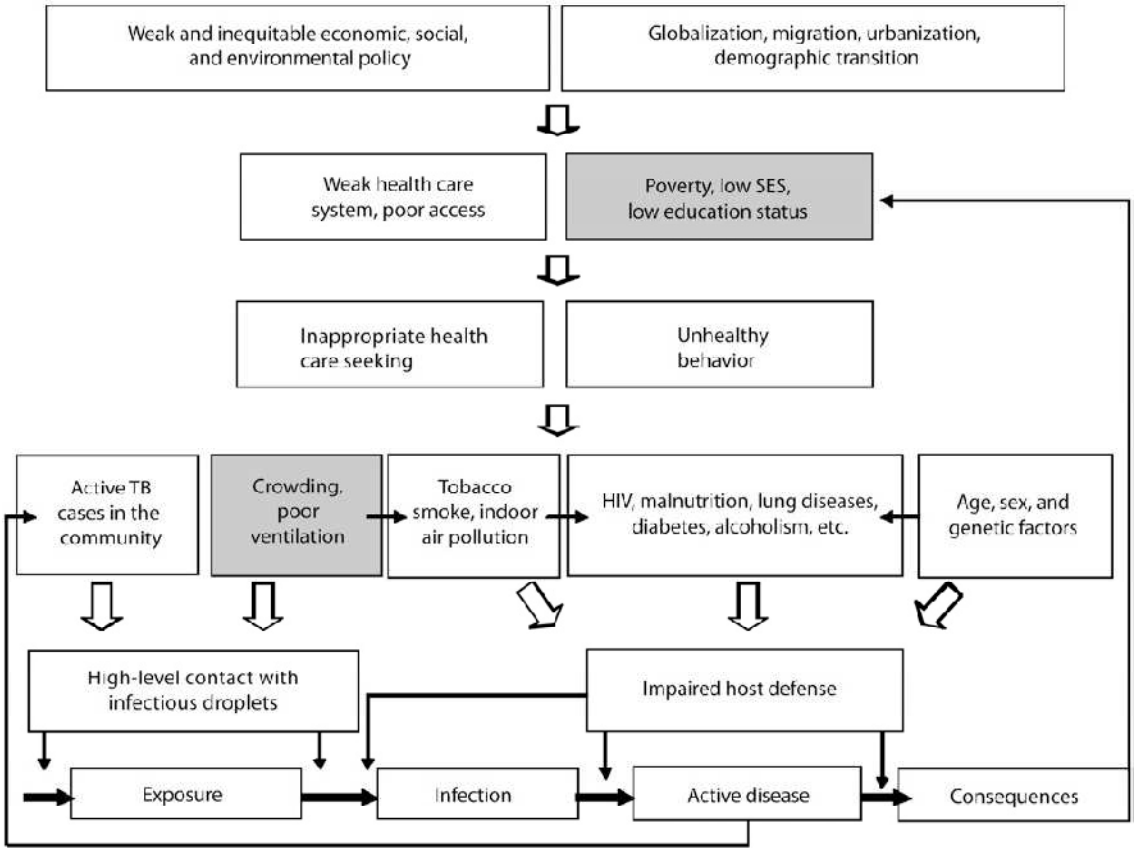
Smoking as a risk factor for TB has been demonstrated by Ghasemian et al (2009) who found out that male smokers were 2.1 times more at risk of developing TB as compared to their non smoking counterparts and this association was dependent on dose and duration. This could be attributed to the fact that smoking causes peribronchial inflammation, fibrosis, thickening of the bronchial wall (intimal) and damage to the alveoli. This in turn reduces ciliary activity leading to poor clearance of inhaled substances. Similar results had been observed by several other prior studies, among them are those conducted by Davies et al (2006) and Chiang et al (2007). Indoor air pollution from solid fuel has also been associated with TB and the mechanism is believed to be similar to that elicited by tobacco smoking. From the studies carried out most show that it increases the risk but the risk is not statistically significant, Lin et al (2007).

HIV has been documented as a risk factor for TB in several studies such as those carried out by Crampin et al (2004) and Lienhardt et al (2005). These studies revealed that increased incidence of TB were observed among the HIV positive as compared to those who were HIV negative. The reason for the trend is reduced CD4 count and increased viral load. In Kenya, HIV is the major cause of TB epidemic. Since 2005, HIV has contributed significantly to increased proportion of smear negative TB that has surpassed notified smears positive TB, DLTLD MOPHS, (2009).

Altitude has also been associated with TB where Mansoer et al (1999) and Tanrikulu et al (2008) observed reduced incidence of TB with increased altitude. To be precise, Tanrikulu found out that the incidence of TB in low altitude was 3.28 times higher than in high altitude.

2.3 1 Conceptual framework

The TB conceptual framework was adapted from the figure below that was developed by Jaramillo et al (2009)



Source: Jaramillo et al (2009) c.f. Hargreaves et al (2011)

Figure 1-TB Conceptual framework.

2.3.2 Operational plan

From the conceptual framework developed by Jaramillo et al (2009) we were able to come up with an operational plan for TB determinants used in the study. The roles of these determinants in TB development are either direct or indirect. HIV and indoor air pollution are considered to have a direct influence on TB since they interfere in one way or another with the normal functioning of the body, through reduction of CD4 count and ciliary activities respectively. The independent factors act indirectly and sometimes contribute to HIV and indoor air pollution leading to active TB.

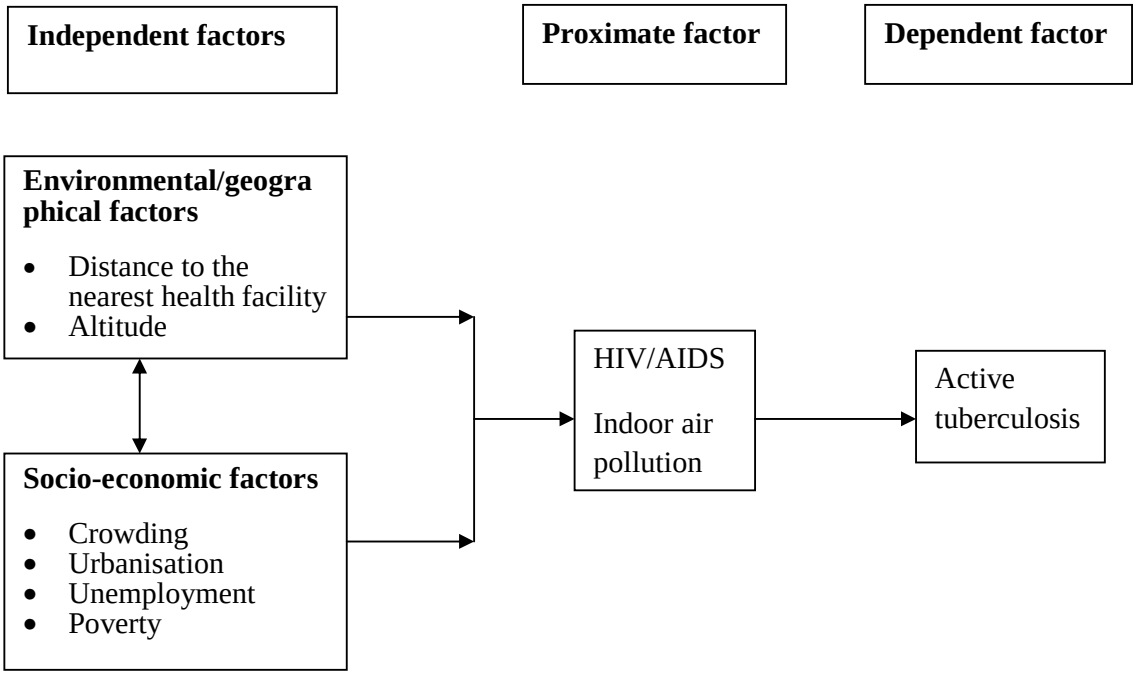


Figure 2-TB Operational plan

3.0 METHODOLOGY

This study used the data submitted to the Division of leprosy tuberculosis and lung diseases every quarter from the DTLCs through their respective PTLCs in Kenya. First we looked at the smoothing techniques then generated maps of tuberculosis at the district level. The maps generated were not of standardised morbidity ratio but rather of smoothed SMR to ensure distinct patterns of TB were observed. Areas with high occurrences of TB (hotspots) were identified in the TB data and the presence of their association with TB determinants investigated.

3.1 Study area, data and population

The study was conducted in Kenya, a country located in East Africa bordered by Uganda, Tanzania, Somalia, Ethiopia and Sudan. The study population included all people at risk of developing TB .Since the study was retrospective, we used annual district TB data from the year 2002 to 2009.This data is usually collected routinely as TB patients are diagnosed and put on treatment. We used all cases of TB i.e. pulmonary sputum positive and negative and extra-pulmonary all forms. Basically the data used incorporated all forms of TB in Kenya found in all age groups. The coordinate information was downloaded from Google maps by using the code: javascript:void(prompt (",gApplication.getMap().getCenter())); This is an application that allows centroid of the specific region to be downloaded once it is identified in the map. The population at risk used was the district projected population as calculated by the Kenya National bureau of statistics for the year 2000 to 2020. In addition, data on HIV was obtained from DLTLD's, poverty information from KIHBS 2005/2006, altitude from www.fallingrain.com on 22nd June 2011 and the determinant factors were used to reveal their significant association with TB.

3.2 Tuberculosis Regional count data

Regional data is a set of counts of diseases available in a set of geographic regions that could be in the form of districts, counties, census tracts and blocks among others. In this study, regional count data partitioned as districts was used to identify areas with high incidences of TB in Kenya .For exploratory spatial data analysis smoothed maps using full Bayesian approach for each year were generated. Then methods used to identify disease clusters and tests to identify disease clustering implemented. The methods for spatial autocorrelation

include Moran's I and the G statistics but the study employed Moran's I because the later can only be used to identify positive spatial autocorrelation and not negative spatial autocorrelation.

Since the research is on spatial temporal techniques, all the newly created districts with TB data available for less than three years were incorporated into their previous districts. This was to ensure uniformity and consistency of results from the districts. Furthermore, some of the newly created district's boundaries are yet to be approved by parliament therefore the coordinates have not yet been geo-referenced into GIS for public use.

3.3 Data preparation

The data used for analysis was sourced from different organizations and or entity. This included data on TB cases per district, the district coordinates (latitude and longitude), HIV prevalence per district, proportion of poverty, Mean district altitude, Mean household size, Employment opportunity, urban areas, proportion of the population that use firewood, proportion of the population per district that are more than 5 kilometres away from the nearest health facility and the projected population per district.

The expected numbers of TB cases were calculated using the population and the observed TB cases.

Let O represent the observed number of TB cases in Kenya in a given year, P the population of that year and R the rate of TB in that year.

Therefore $$R = O / P \qquad (1)$$

And let O_i be the observed number of TB cases per district where $(i = 1, 2.....N)$, E_i be the expected number of TB cases per district and P_i the population of each district.

The expected number of TB cases per district was:

$$E_i = R \times P_i \qquad (2)$$

After calculating the expected number of TB cases per district per year from 2002 to 2009, the data set was saved in excel as comma delimited file ready for analysis.

The comma delimited data set and the Kenya district boundary shape file were imported into R. These two files were then merged to generate one data set that had all the variables of interest including the co-ordinates (See R commands in appendix 1)

Secondly in the same data set, number of columns and rows were indicated in the top most rows then imported into Geoda software. Here, the Kenya district centroid shape file was created from Kenya district boundary shape file. The Kenya centroid shape file was then opened in excels as a dbase file and its centroids copied into the comma delimited file. This file was then converted into a point from ASCII shape file. The shape file was later joined with the Kenya centroid dbase file and the output used again to be merged with the Kenya district boundary shape file. The whole process enabled the merging of the data to the Kenya district boundaries thus creating a polygon file with the required data. This polygon file was then used for analysis. Apart from polygon file, weight files were also created to be used in computing Moran's I statistics and in Bayesian analysis.

However, the data used for space time scan statistics was prepared in a different manner such that three files namely; case file, population file and coordinate files were created. The case file had the district, cases and the time period in years, the population file had the district, population and the time period in years and the finally the coordinate file had the district, latitude and longitude. All the files were saved as comma delimited files and imported into Sat Scan software for spatiotemporal analysis.

3.4 Statistical methods

3.4.1 Smoothing models

First standardized morbidity ratio was calculated using Clayton and Kaldor (1987) approach. This is an indirect method of standardization where the observed number of event in a study population is compared to the expected number of the event in the same study population if the event were the same as that of the standard population. In our case the event was TB cases or occurrence of TB. The ratio is therefore the observed number of TB cases to expected number of TB cases.

So that

$$SMR_i = O_i / E_i \qquad (3)$$

The SMR_i was used to indicate the relative excess or decrement of TB cases in each district with respect to the standard TB rate in Kenya as a country for each year.

When modelling count data, the approach is to assume that the number of cases follow a Poisson distribution. A Poisson distribution models assumes that the mean and the variance are equal but this is not always the case. In most situations the data is "over-dispersed" meaning that the variance is higher than the mean. As a result the model needs to be expanded to cater for the over-dispersion. One way of doing this is to use a negative-binomial distribution.

When negative binomial distribution is used, a random effect following a gamma distribution for each region can be considered to produce what is called the Poisson-Gamma model. The Poisson-Gamma model is structured in two levels as shown below.

$$O_i \mid \theta_i, E_i \sim Po\left(\theta_i E_i\right), \tag{4}$$

$$\theta_i \sim Ga\left(v, \alpha\right) \tag{5}$$

From above θ_i is the relative risk which is considered as a random variable drawn from a Gamma distribution (scale parameter α and shape parameter v) with mean $\left(v/\alpha\right)$ and variance $\left(v/\alpha^2\right)$. In the first level, the distribution O_i is conditioned on θ_i. But in the second level, O_i is a negative binomial with size parameter v and probability $\frac{\alpha}{\alpha + E_i}$. Also the posterior distribution of θ_i has a Gamma distribution with shape parameter $v + O_i$ and scale parameter $\alpha + E_i$. This means that the information after observing the data has been updated and therefore the posterior expectation of θ_i is:

$$E\left[\theta_i \mid O_i, E_i\right] = \frac{v + O_i}{\alpha + E_i} \tag{6}$$

Therefore the shrinkage estimator can be expressed as

$$E[\theta_i \mid O_i, E_i] = \frac{E_i}{\alpha + E_i} SMR_i + \left(1 - \frac{E_i}{\alpha + E_i}\right)\frac{v}{\alpha} \tag{7}$$

The output in equation (6) and (7) is usually affected by O_i especially in low populated areas where a small variation of observation can produce a big change in SMR_i. In addition, given that the value for v and α are the same for all the regions, the information is borrowed from

everywhere to construct the posterior estimates but this can be modified to consider a different set of neighbours.

Clayton and Kaldor (1987) developed an approach that used Empirical Bayes estimation. The estimation of v and α in a negative binomial distribution can be done using the maximum likelihood function. The log likelihood based on the distribution of the O_i is:

$$L(\alpha, v) = \sum_i \left[\log \frac{\Gamma(O_i + v)}{\Gamma(v)} + v \log(\alpha) - (O_i + v) \log(E_i + \alpha) \right] \quad (8)$$

Setting the first derivative in the above equation with respect to α and v to yield 0, the equation becomes;

$$\frac{\hat{v}}{\hat{\alpha}} = \frac{1}{N} \sum_i \frac{O_i + \hat{v}}{E_i + \hat{\alpha}} = \frac{1}{N} \sum_i \hat{\theta}_i \quad (9)$$

and

$$\sum_{i=1}^{N} \sum_{j=0}^{O_i - 1} \frac{1}{v + j} + N \log(\alpha) - \sum_{i=1}^{N} \log(E_i + \alpha) = 0, \quad (10)$$

where $\sum_{j=0}^{-1} 1/(v + j) = 0$

Equation (9) and (10) can be solved using standard iterative procedures thus yielding the empirical Bayes estimates $\left(\hat{\theta}_i\right)$ as shown in the equation below.

$$\frac{\hat{v}}{\hat{\alpha}^2} = \frac{1}{N-1} \sum_i \left(1 + \frac{\hat{\alpha}}{E_i} \right) \left(\hat{\theta}_i - \hat{v} / \hat{\alpha} \right)^2 \quad (11)$$

where N-1 is a Pearsonian chi-square based on mean and variance above.

It is also possible to allow the model for the distribution of (θ_i) to extend for covariates such as (z_i). Therefore the estimates of the districts with few observed and expected cases of TB will not be drawn to the global relative mean of TB but rather to the estimated degree of the

covariate being used. This is done by allowing distinct values, (α_i), for the scale parameters of the distributions of each θ_i, and adopting the log-linear model shown below:

$$E(\theta_i) = \frac{\nu}{\alpha_i} = \exp\left(z_i^T \phi\right) \tag{12}$$

In Empirical Bayes technique developed by Marshall (1991, c.f. Bivand et al 2008) the relative risk θ_i was assumed to have a common mean μ and variance σ^2 without being specific on any distribution. Using the method of moments, the shrinkage estimator proposed was:

$$\hat{\theta}_i = \hat{\mu} + C_i\left(SMR_i - \hat{\mu}\right) = (1 - C_i)\hat{\mu} + C_i SMR_i \tag{13}$$

where

$$\hat{\mu} = \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} E_i} \tag{14}$$

and

$$C_i = \frac{s^2 - \hat{\mu}/\overline{E}}{s^2 - \hat{\mu}/\overline{E} + \hat{\mu}/E_i} \tag{15}$$

When mapping disease incidence using conventional methods, one may not get the true spatial pattern because this methods are affected by random variation as a result of differing population in different regions. This leads to reduced statistical power when the cases are assigned to different subgroups such as district. This may further lead to incorrect interpretation as true epidemiologic variation. Bayesian methods models random variation and true variation therefore it can be used to avoid the problem. In addition, Bayesian models provide shrinkage and spatial smoothing thus allowing distinct pattern of disease to be observed.

It is also evident that when hierarchical Bayesian models are computed, Waller and Gotway (2004), layers could be created such that each of these layers account for a specific source of variation. This means that covariates can be modelled as the model borrows information from

the neighbours (smoothing). This allows achievement of stable local rates without losing geographic resolution.

One of the drawbacks of empirical Bayes estimate is its inability to take into account a measure of uncertainty of the estimates, Maiti (1998). Spatial rate smoother on the other hand represent regional averages therefore cannot be used to identify outliers since it's not specific to individual location. Head banging smoother and spatial filters (adaptive k smoothing) are not appropriate because they do not account for variance instability in rates. Kernel smoother on the other hand is good for point data and not region based data. For this reasons we used Bayesian methods.

In the Bayesian model analysis for mapping, the relative risk estimate was obtained through a measure of central tendency of the posterior distribution and the median. Bayesian Poisson Gamma formulations were used where v and α were assigned vague gamma priors to ensure that very little information was introduced into the model. Monte Carlo method with one Markov chain was used to obtain the posterior distribution with Gibbs sampling algorithm. Gibbs sampling was used to overcome the problem of high dimensional integration in the posterior distribution. This allowed us to use multi-path iterative sampling to check the convergence of Gibbs sampling. The Gibbs sampling had 200,000 iterations, a burn in period of 10,000 iterations and a thin rate of 100. This implied that the first 10,000 iterations were discarded to eliminate the effects of the initial choices and one in every hundred of the remaining iteration was stored. It is important to note that a package called R2WinBUGS in R software was used to call WinBUGS software using the scripting language and then reads the output log file from it.

3.4.2 Detection of disease clusters

According to Tobler's law, observations closer to each other have related values or are more alike than those further apart. Therefore spatial autocorrelation is a phenomenon that is used to show whether near events are more alike than those further apart. In other words, it shows the relationship between an event in a given location and the same event at other locations.

3.4.2.1 Homogeneity of relative risk

Before any cluster detection, a heterogeneity test was computed in order to find out if there is an actual difference in TB relative risk in the different districts in Kenya. This was done

using a chi-square test to find out if there is a significant difference between the observed and expected number of TB cases in each district. The statistics is defined as below;

$$x^2 = \sum_{i=1}^{N} \frac{\left(O_i - \theta E_i\right)^2}{\theta E_i} \qquad (16)$$

where θ is the global $SMR = \sum_i O_i / \sum_i E_i$

3.4.2.2 K nearest neighbour

While generating weights used in computing spatial autocorrelation, the modality of identifying the neighbours was specified. The study used k nearest neighbour as a method of generating weight. This is a method or approach that is used to assess neighbour relationship regarding a particular event in a certain region. In other words, the event is assessed within the spatial context of a fixed number of its closest neighbour. For example if k is 4, then the four closest neighbours to the target region will be included in the computation of the event.

This method is advantageous in the sense that it ensures that there is some neighbour for each target region and the number of neighbours is fixed. K is the number of neighbours and it should be specified.

3.4.2.3 Moran's I

For this study both global and local Moran's I were utilized for detecting clustering and the presence of clusters in the districts respectively. The global Moran's I was used to provide an insight on the spatial similarity of TB relative risk in the neighbouring districts in the entire study area and was expressed as shown below-:

$$I = \left(\frac{1}{s^2}\right) \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(Y_i - \bar{Y}\right)\left(Y_j - \bar{Y}\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \qquad (17)$$

where $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

and

$$s^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$$

Y_i is the value for a variable in the ith observation $\overline{Y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.

After computing the global Moran's I with Empirical Bayes standardization, randomization procedure was used to recalculate the statistics in order to validate their significance. To generate reference distribution, 999 permutations were applied with a significance filter of $p \leq 0.05$. The p value generated compared the observed statistics to the generated distribution through randomization i.e. it determines how likely it is to observe the actual spatial distribution if the actual values are randomly reshuffled over space at a certain number of permutation.

In Beijing, Jia et al (2008) used global Moran's I with a z score and Getis's G statistics to identify hotspot districts for TB. Likewise, this study employed similar statistics to identify TB hotspot districts in Kenya. Since Getis's G statistics does not show negative spatial autocorrelation, it was not computed instead global Moran's I was used. Apart from that LISA inform of cluster maps were generated to visualize areas of both positive and negative spatial autocorrelation.

Local Moran's I a local indicator of spatial autocorrelation was computed to show the local measure of similarity in each region's value and those of their neighbours. The measure detected individual clusters such that if we have region i, then the local Moran's I was computed as shown below-:

$$I_i = \left(Y_i - \overline{Y}\right)\sum_{j=1}^{N} w_{ij}\left(Y_j - \overline{Y}\right) \qquad (18)$$

To detect local clusters in the TB data set, LISA with empirical Bayes rate was also computed where cluster maps were generated to show the relationship among different districts with their neighbours. These maps were used to identify local spatial clusters and

spatial outliers of the geo-referenced TB data. The relationships were grouped into four categories;

High-high: districts with high relative risk of TB surrounded by neighbours with similar values.

Low- low: districts with low relative risk of TB surrounded by neighbours with similar values

High-low: districts that experience high relative risk of TB surrounded by neighbours with low TB risk.

Low-high: districts with low relative risk of TB surrounded by neighbours with high risk of TB.

It is worth noting that since our population differs so much from one district to another, the Moran's I computed were standardised using Empirical Bayes technique as outlined below.

3.4.2.4 To adjust for the population using EB technique

Assun¸c˜ao and Reis (1999) developed a Moran's I that uses Empirical Bayes rate to shrink extreme rates with small population at risk towards the entire study area's rate. In this technique a new index is generated using deviation of estimated marginal mean standardized by the estimate of its standard deviation.

Let $\theta_1.....\theta_N$ be unknown rate of TB in the different districts, and O_i observed TB cases per year with a Poisson distribution conditional mean of $E(O_i \mid \theta_i) = P_i\theta_i$ and the estimated rate R_i has conditional mean $E(R_i \mid \theta_i) = \theta_i$ and variance $\text{var}(R_i \mid \theta_i) = \theta_i / P_i$.

From above, if we denote priori expectation and variance as β and α respectively then the marginal expectation of R_i is β and marginal variance $\alpha + \beta/P_i$. This shows that the variance is different and it increases as the population decreases.

Marshall (1991 c.f. Assun¸c˜ao and Reis 1999) proposed moment estimators for priori expectation and variance as; $b = O/P$ and $a = s^2 - b/(N)$

where $s^2 = \sum P_i (R_i - b)^2 / P$

and N is the number of areas/regions,

O is the observed event and

P is the population at risk

Then, the marginal expectation and variance of R_i were estimated by b and vi respectively where $v_i = a + b/P_i$

The new index proposed therefore was:

$$z_i = \frac{R_i - b}{\sqrt{v_i}} \qquad (19)$$

and the Empirical Bayes Index was defined as

$$EBI = \frac{N}{\sum w_{ij}} \frac{\sum w_{ij} z_i z_j}{\sum \left(z_i - \bar{z}\right)^2} \qquad (20)$$

Before computing LISA with Empirical Bayes standardization, the Geoda software was set at 999 permutations with a significance filter of $p \leq 0.05$. This was to ensure that the cluster maps generated showed areas with significant difference at 5% level of significance.

For interpretation of spatial association, the indicators results were used to suggest spatial autocorrelation with 1 and -1 depicting strong positive and negative spatial autocorrelation respectively. A zero value indicated a spatial random pattern.

3.4.3 Space time scan statistics

A spatial scan statistics developed by Kulldorff and Nagarwalla (1995 c.f. Kulldorff 2010) was used to identify and detect clusters in a particular district together with their statistical significance. This statistics works by imposing a circular or elliptic window on the map and allowing its centre to move over the study region. The window can take any predefined shape and its size may be allowed to vary as it moves and scans the study region, Kulldorff (2010).For any centre position, the radius of the circle is able to change so that it takes any value between zero and some upper limit set by the user. The upper limit set by the user can be a percent of the population used in the analysis, a percentage of population defined in the

maximum circle file or geographic size. The maximum upper limit for spatial analysis that can be used is 50% of the total population at risk, otherwise if the cluster size is large, it would indicate areas of low rates outside the circle rather than areas of high rates within the circle, Kulldorff (2010). Having 50% of the total population at risk is the preferred value since SaTScan identifies both small and large clusters without any bias on cluster size.

Each of these circle generated includes different sets of neighbouring districts so long as they are within the circle and is a potential cluster for TB incidence. For each circle generated, spatial scan statistics calculates its likelihood for the TB cases observed inside and outside the circle. From the circles computed, the one with the maximum likelihood is considered the most likely cluster meaning that it could not have occurred by chance. This method tests the null hypothesis that the risk of TB is the same in all the districts in Kenya.

Since we are using regional count data, a retrospective Space time scan statistic was computed on the TB data from 2002 to 2009 using SaTScan software. The likelihood ratio based on Poisson distribution was used. Therefore the likelihood ratio for a Poisson assumption for a specific window was defined as:

$$\max\left(\frac{O_{in}}{E_{in}}\right)^{O_{in}}\left(\frac{O_{out}}{E_{out}}\right)^{O_{out}} I() \qquad (21)$$

where O in is observed cases of TB in the defining window, $E\ in$ is the expected cases in the defined window, O out and $E\ out$ are the observed and expected cases outside the window of interest respectively. I is an indicator function which is equal to 1 when the circle has more TB cases than expected under the null hypothesis and 0 otherwise.

For analysis, a discrete Poisson model with a cylindrical window where a circular geographic base and a height corresponding to time was imposed to move over the study region in time, so that for each possible area and size, it also visited each possible time period. For spatial 50% of the population at risk was the specified maximum cluster size while for temporal 50% of the study period was also specified as the max cluster size. Standard Monte Carlo simulation with a significance level of 5% was used for inference, to test the null hypothesis of spatial randomness. It was rejected when the p value assigned to the detected cluster was found to be significant. No spatial, temporal and or space time adjustments were made and the number of Monte Carlo replications was set at 999. As a criterion for reporting secondary

clusters (areas with high rates of TB but with lower likelihood than the most likely cluster), no geographic overlap was allowed.

As stated above, the most appropriate space time scan statistics for our data is a discrete Poisson model because the data is in count form and the population for each district is provided. Other models that space time scan statistics apply are; Bernoulli model which requires cases and control data, Ordinal model that requires ordered data in categories, multinomial model where case data that is grouped into specific categories is needed and space time permutation model that needs only case data but then it cannot be used where population growth in different regions of the study area is not proportional and when the study period is long, Kulldorff (2010). Furthermore Exponential model requires survival data and finally normal model utilizes continuous data.

3.4.4 Spatial modelling

To determine the association of TB and the risk factors, a Bayesian approach was applied where Conditional Autoregressive Model (CAR) was specified. This model depends on the conditional distribution of spatial error terms and explains part of the variability of the relative risk. The CAR specification for a set of random variables $\{v_i\}_{i=1}^{N} = 1$ can be defined as below:

$$v_i \mid v_{-i} \sim N\left(\sum_{j \neq i} \frac{w_{ij} v_j}{\sum_j w_{ij}}, \sigma_v^2 / \sum_j w_{ij}\right) \qquad (22)$$

Where $v_i \mid v_{-i}$ is the vector of all error terms minus the error term itself. w_{ij} is the weight that measures the relationship between region i and j while σ_v^2 shows the conditional variance of the CAR model specified.

3.4.4.1 The Model

Before computing the model we needed to specify the model using the BUGS (Bayesian inference using Gibbs sampler) language, the TB data, spatial data describing the neighbourhood structure and initial values of the parameters (see R commands appendix 1)

$$O_i \sim poisson(\mu_i)$$

and $\mu_i = \theta E_i$

therefore

$$\log(\theta_i) = \alpha + \beta x_i + u_i + v_i$$

v_i is the CAR specification , u_i is the non spatial random effect and x_i is the explanatory variable.

The weight file used in this model was queen continuity but it was converted into an nb2WB in R. This allowed us to set all the weights for adjacent neighbours to 1 and 0 otherwise. During the analysis, MCMC simulation method was used and 100,000 iterations specified where the first 10,000 were discarded leaving 90000 and each 100th sample was stored. This sample was then used to assess convergence in the trace plots generated. By including the covariates, we were able to assess and remove the effects that occur as a result of confounding or risk factors. To know the effects of the covariate, we looked at the posterior density plot of β. If the posterior mean or median was greater than zero, the coefficient of the covariate was considered to be positive and when the 95% credible interval did not include zero, then it was interpreted that the coefficient is statistically significant. This means that there exists a positive association between TB and the risk factor.

To diagnose convergence, several chains in the trace plot showing Markov chain Monte Carlo (MCMC) simulations output were compared and if they appeared to concentrate around the posterior mean or median of the 2002 TB data, then it was considered that convergence had taken place.

4.0 RESULTS

4.1 Bayesian smoothed Maps of TB

The Bayesian smoothed maps labelled figure 3 to figure 10 below, shows the posterior mean and median of the TB relative risk in Kenyan districts from the year 2002 to 2009. These maps were used for exploratory spatial data analysis to show the spatial patterns of TB occurrence in Kenya and identify high risk districts.

In 2002, the districts with high relative risk of TB were: Bondo, Embu, Isiolo, Mandera, Marsabit, Meru south, Moyale, Nairobi, Rachuonyo, Homabay, Kisumu, Mombasa, Kitui and Nyando. In 2003, all the districts that had had high relative risk of TB were maintained apart from Marsabit. In addition, Nakuru and Suba also recorded a significant relative risk during this year.

In 2004, Bondo, Embu, Homabay, Kitui, Isiolo, Meru south, Moyale, Migori, Mombasa, Nairobi, Nakuru, Suba, Rachuonyo, Turkana, Uasin Gishu , west pokot, Kisumu and Nyando experienced high TB occurrence while in 2005 all these districts maintained their high rates in addition to Mandera and Siaya but excluding Turkana and west pokot.

The year 2006 didn't depict a very different spatial pattern since similar districts recorded high occurrences of TB. Kitui, Nakuru and Meru south relative risks declined this year and therefore were not among the TB hotspot identified. In 2007, Busia and Kuria emerged as hotspot since these two districts had not recorded high TB occurrences compared to the country's rate. The other districts especially those around the lake region maintained high relative risk of TB.

In 2008, Bondo, Homabay, Isiolo, Kisumu, Migori, Moyale, Mombasa, Nairobi, Nyando, Rachuonyo, Siaya, Suba, Turkana and Malindi had high rates of TB. It is worth noting that Malindi emerged as hotspot in 2008. In 2009, Tharaka and Thika were identified as emerging hotspot areas but Bondo, Homabay, Isiolo, Kisumu, Migori, Moyale, Nairobi, Rachuonyo, Siaya, Mombasa, Malindi, Turkana, Uasin Gishu and Suba continued to display their high TB occurrence.

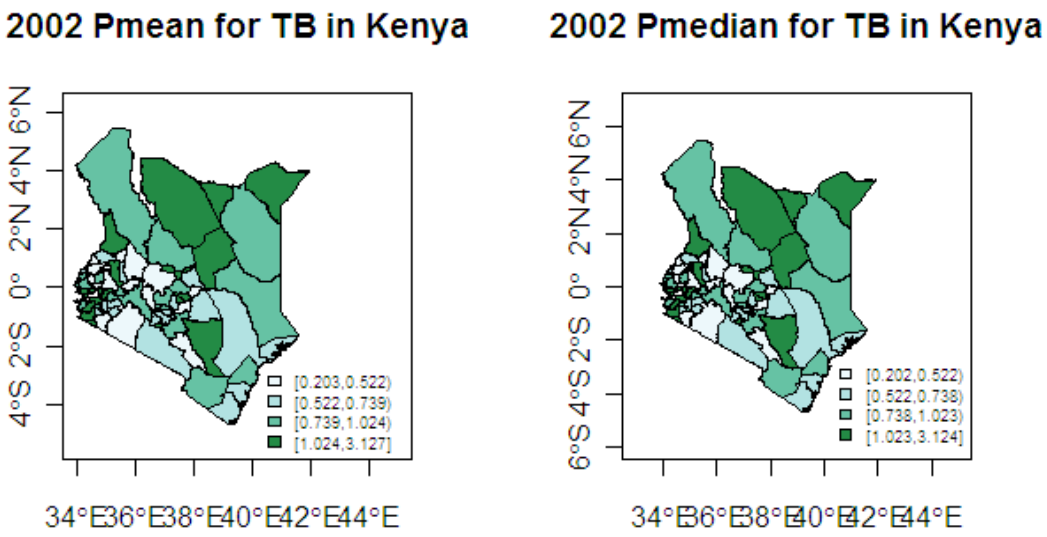


Figure 3-Maps of posterior mean and median relative risk of TB for the year 2002

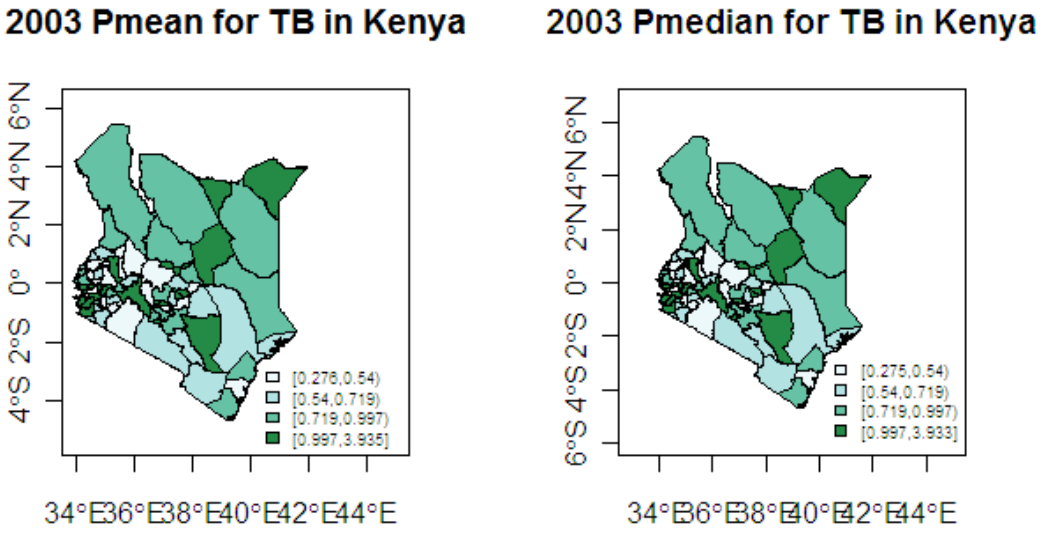


Figure 4-Maps of posterior mean and median relative risk of TB for the year 2003

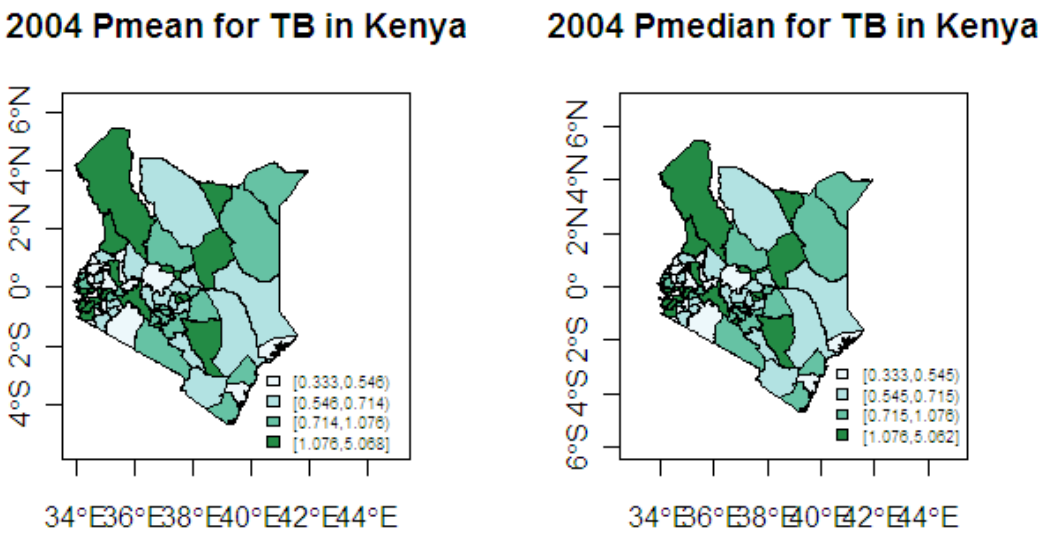


Figure 5- Maps of posterior mean and median relative risk of TB for the year 2004

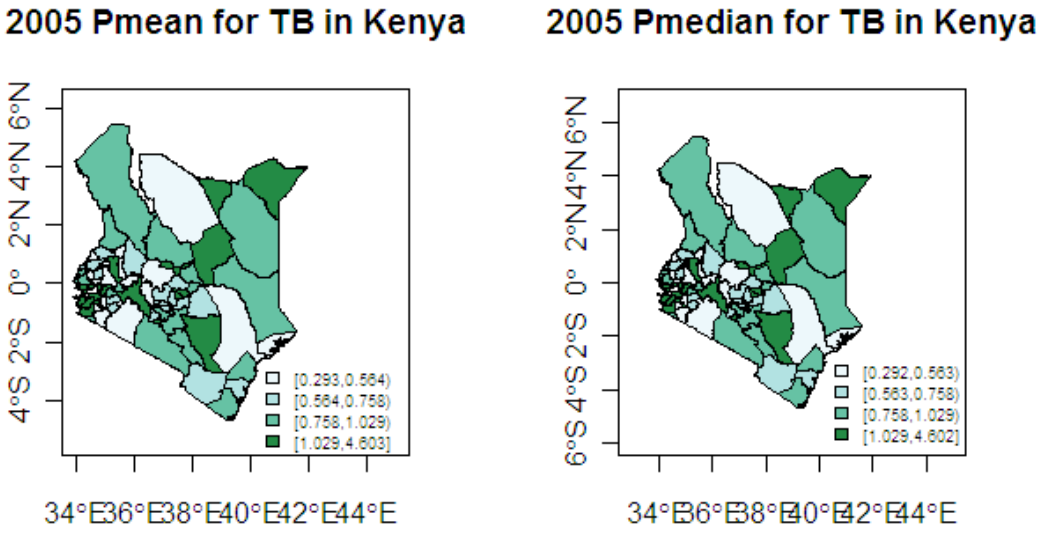


Figure 6-Maps of posterior mean and median relative risk of TB for the year 2005

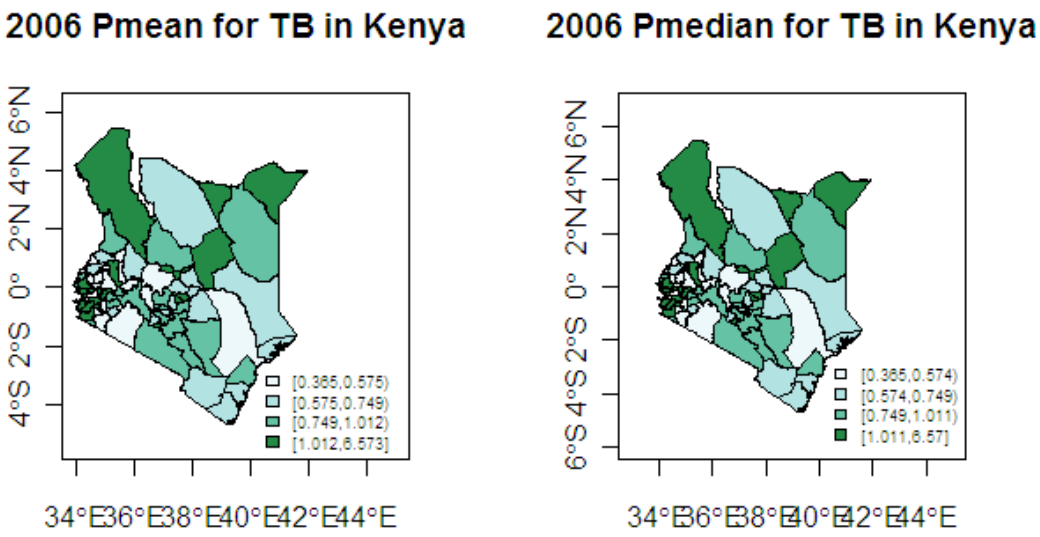


Figure 7-Maps of posterior mean and median relative risk of TB for the year 2006

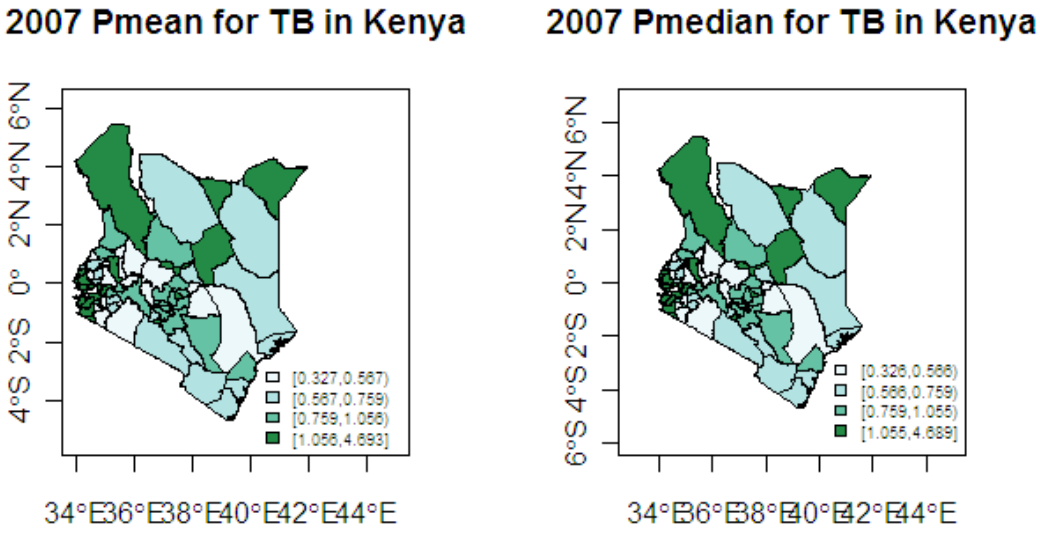


Figure 8-Maps of posterior mean and median relative risk of TB for the year 2007

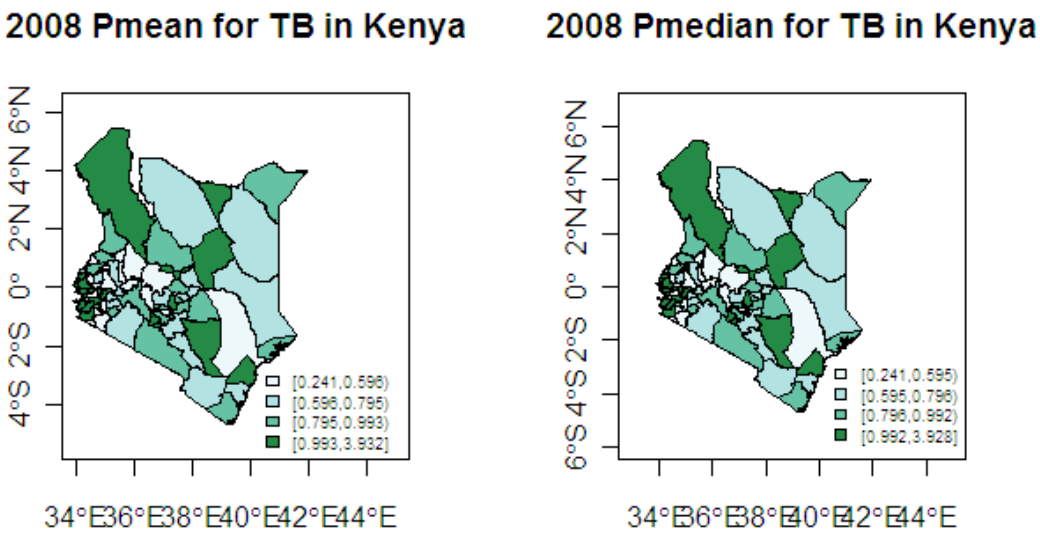


Figure 9-Maps of posterior mean and median relative risk of TB for the year 2008

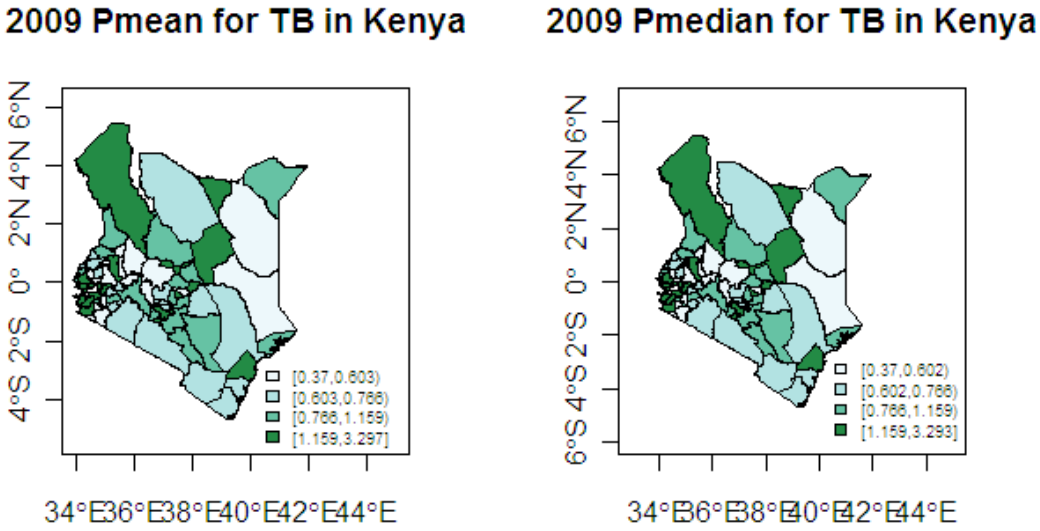


Figure 10-Maps of posterior mean and median relative risk of TB for the year 2009

4.2 Testing homogeneity of the relative risk

A chi-square test was computed on TB data on a yearly basis to test for homogeneity before proceeding to find out more on clustering and presence of clusters. The number of simulation used was set at 999 and results are shown in table 1 below:

All the results suggest that there is over dispersion which is significant at 5% level of significance. This is because the p value is less than 0.05 suggesting strong evidence against homogeneity.

Table 1-Results of a chi-square test for homogeneity of TB data from 2002 to 2009

Year	Statistics	P.value
2002	37895.37	0.001
2003	39823.35	0.001
2004	40298.76	0.001
2005	38765.44	0.001
2006	40855.12	0.001
2007	33966.95	0.001
2008	32771.18	0.001
2009	31823.29	0.001

4.3 Spatial autocorrelation

Spatial autocorrelation techniques used to identify the existence of spatial association either in the entire study region or the individual regions with their neighbours was computed in Geoda software. The weight file computed had four as the specified number of nearest neighbours.

4.3.1 Global Moran's I

The results of global Moran's I are shown in table 2 below.

These results suggest that there is positive spatial association on the TB data as the global Moran's I is greater than zero. From 2002 to 2006 the positive spatial similarity is not significant while from 2007 to 2009 the spatial similarity is significant at 95% confidence interval since the p. values generated after randomization are less than 0.05. This means that the relative risks of TB in near regions are more alike than those further away.

Table 2-Results of Global Moran's I for TB data from 2002 to 2009.

Year	Moran's I standard deviate	Moran's I statistics	P value	Expectation	Mean
2002	0.0715	0.0898	0.093	-0.0147	-0.0135
2003	0.0720	0.1122	0.055	-0.0147	-0.0159
2004	0.0667	0.0619	0.129	-0.0147	-0.0168
2005	0.0685	0.1022	0.067	-0.0147	-0.0140
2006	0.0559	0.0781	0.068	-0.0147	-0.0099
2007	0.0661	0.1217	0.029	-0.0147	-0.0178
2008	0.0707	0.1540	0.016	-0.0147	-0.0190
2009	0.0710	0.1164	0.044	-0.0147	-0.0140

4.3.2 Local Indicators of Spatial Autocorrelation (LISA)

Below are cluster maps labelled figures 11 to figure 18 generated to identify the local clusters of TB in the Kenyan districts for each year from 2002 to 2009. Spatial clusters in terms of hot spots and cold spots were defined. The spatial outliers were also identified because they have an impact in epidemiology and disease surveillance.

In 2002, hotspot districts identified were Mandera, Moyale and Marsabit. These areas had high TB values surrounded by neighbours with high TB occurrence. Garissa, Kakamega,

Trans Nzoia had low TB values surrounded by neighbours with low TB occurrence. Wajir had low values of TB surrounded by neighbours with high TB value. West pokot and Uasin Gishu had high TB relative risk surrounded by neighbours with low risk.

In 2003, Suba emerged as one of the districts with high TB occurrence surrounded by neighbours with high values of TB. Bungoma and Mt Elgon were added as cold spots. The spatial outliers remained i.e. Uasin Gishu, West pokot and Wajir. In 2004, Mandera was the only hotspot revealed, the cold spots remained but Marsabit was added among the spatial outliers that were surrounded by neighbours with high TB incidence. Clusters map for the year 2005 was similar to that of 2004 with only one additional hot spot district called Suba.

In 2006, Mandera remained as a hotspot with another emerging hotspot in Bondo. The emerging cold spot were West pokot, Samburu and Teso. Marsabit and Wajir remained as spatial outliers surrounded by neighbours with high occurrence of TB. Uasin Gishu still continued to be surrounded by neighbours with low TB cases.

Districts with high TB incidence surrounded by neighbours with high TB occurrence in 2007 were Suba and Mandera. Trans Nzoia, Bungoma, Kakamega and Garissa remained as cold spots. Marsabit and Wajir maintained the same spatial pattern while West pokot and Uasin Gishu had high values surrounded by neighbours with low values of TB.

The only hotspot revealed in 2008 was Suba but Trans Nzoia emerged as an area with high TB occurrence surrounded by neighbours with low TB values. Cold spots for this year were Baringo, Bomet, Keiyo, Kakamega, west pokot and Uasin Gishu. Mandera was no longer an outlier but an area surrounded by high values of TB.

In 2009, Bondo and Suba were the high TB areas surrounded by neighbours with high TB occurrence. Kakamega and Trans Nzoia remained as cold spots. Uasin Gishu again was identified as an area with high TB relative risk surrounded by neighbours with low TB values.

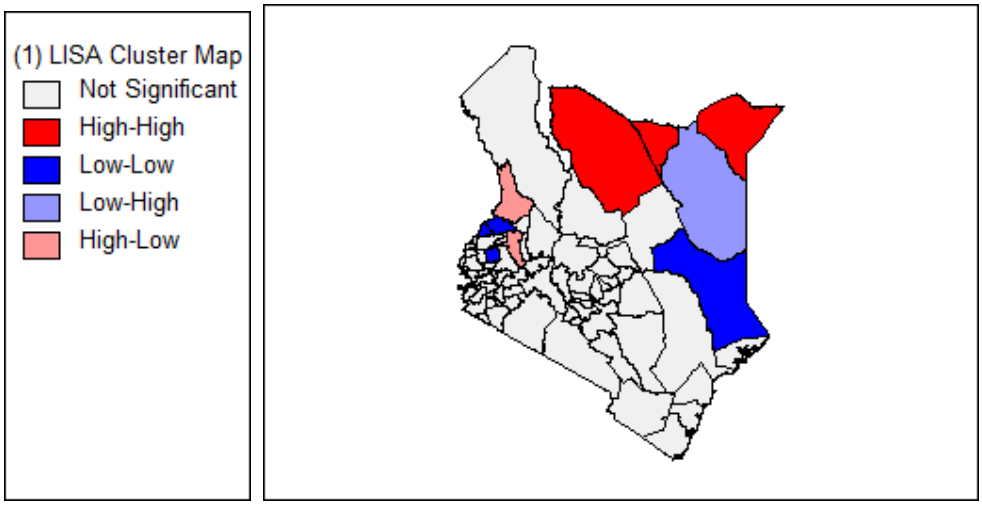


Figure 11-Empirical Bayes standardized Cluster map of TB for the year 2002

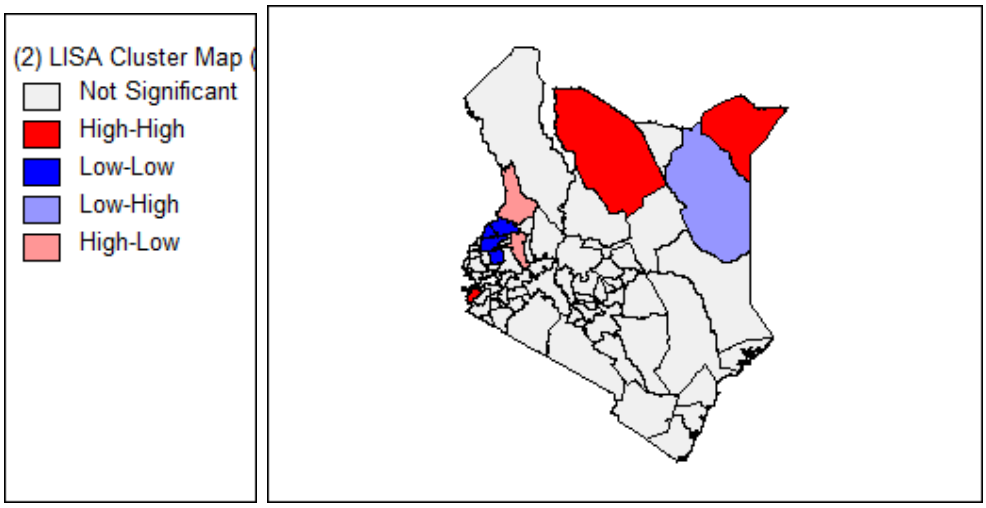


Figure 12-Empirical Bayes standardized cluster map of TB for the year 2003

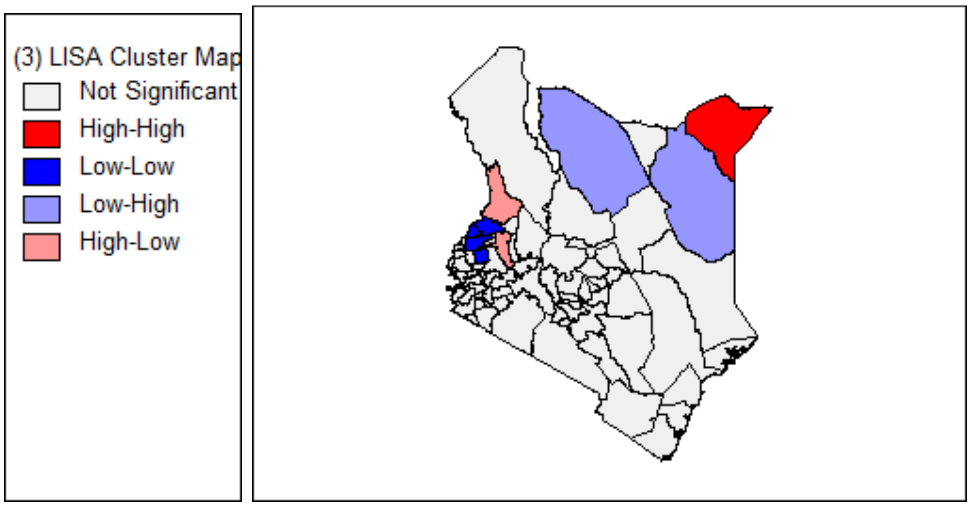


Figure 13-Empirical Bayes standardized cluster map of TB for the year 2004

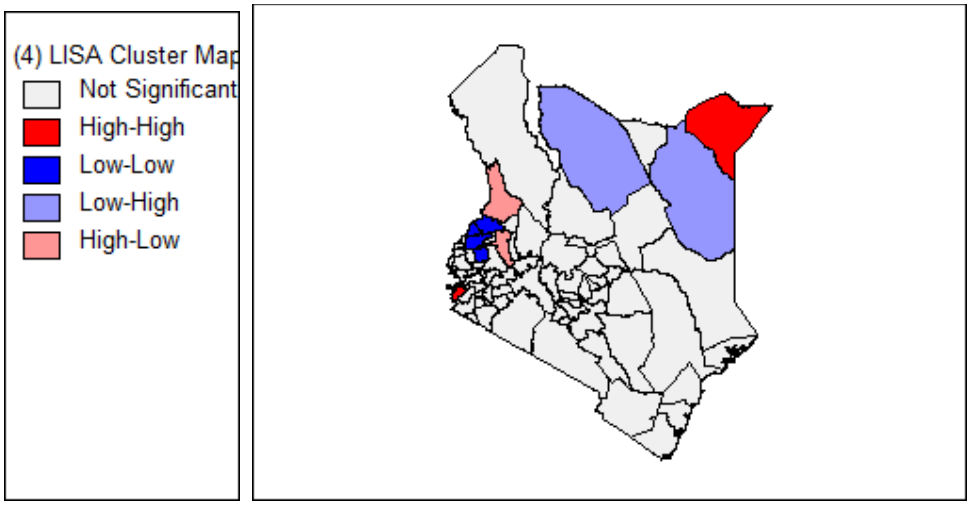


Figure 14-Empirical Bayes standardized cluster map of TB for the year 2005

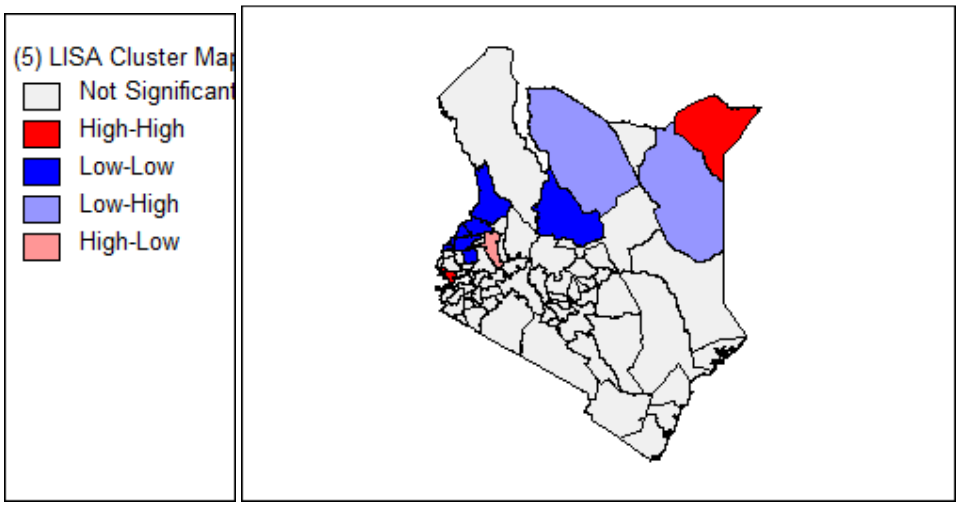


Figure 15-Empirical Bayes standardized cluster map of TB for the year 2006

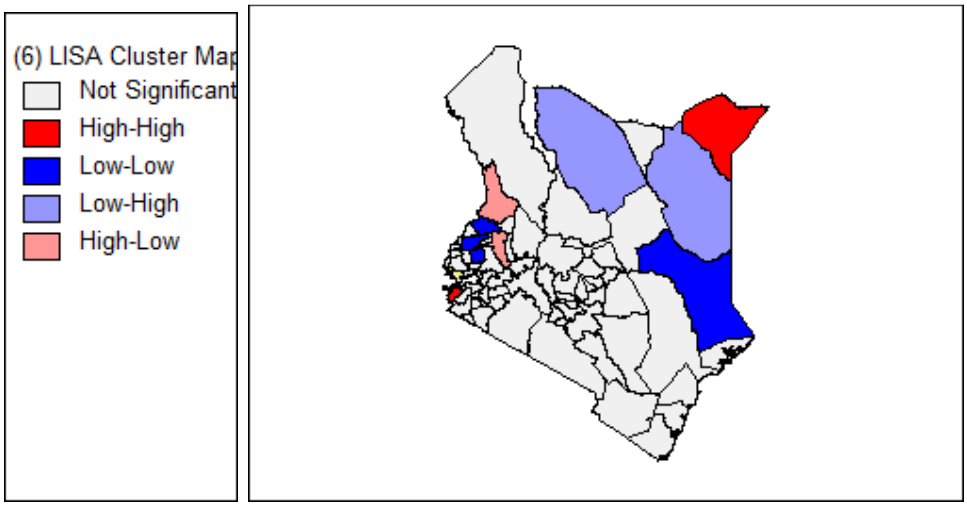


Figure 16-Empirical Bayes standardized cluster map of TB for the year 2007

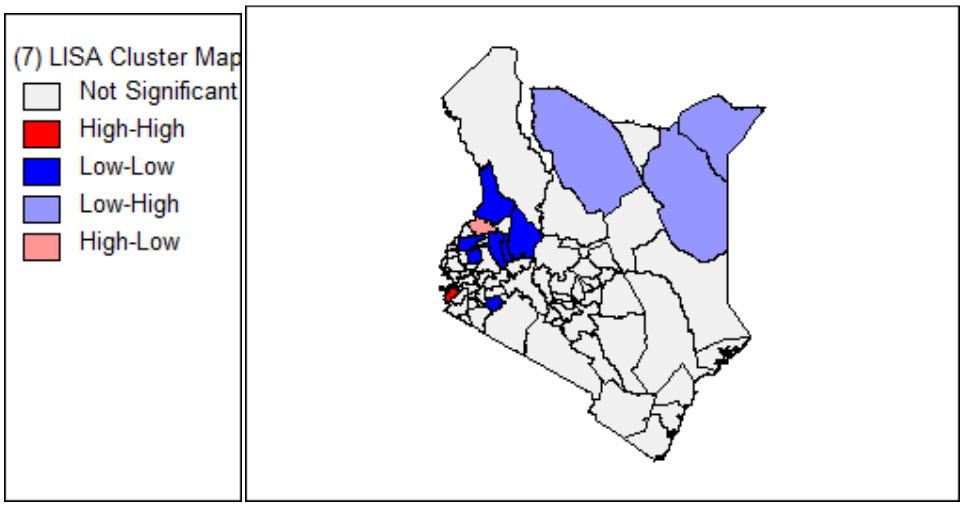


Figure 17-Empirical Bayes standardized cluster map of TB for the year 2008

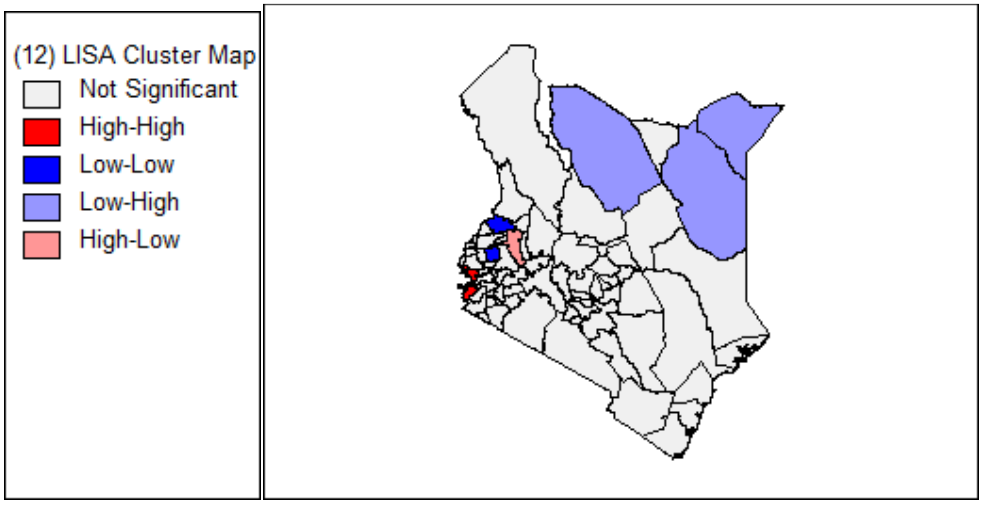


Figure 18-Empirical Bayes standardized cluster map of TB for the year 2009

4.4 Space time scan statistics

In space time analysis, we tested the areas with high TB occurrence from the 69 districts in Kenya starting from first of January 2002 to thirty first of December 2009. Remember that the spatial window was set at 50% therefore the maximum number of year computed for a cluster is four years. These results are shown in Table 3 below.

The most likely cluster identified that would not have occurred by chance was Nairobi district with a statistical significance cluster ($p<0.001$) from 2003 to 2006. The observed cases were 77190 while the expected cases were 32414. It had the highest log likelihood ratio thus classified as the most likely cluster. The relative risk is as shown in Table 3.

The other clusters identified as secondary which were statistically significant ($p<0.001$) from 2005 to 2008 were Suba, Bondo, Homabay, Migori, Siaya, Rachuonyo and Kisumu. Their observed TB cases were 63844 with 35694 as the expected number of cases. In addition several other secondary clusters that were statistically significant ($p<0.001$) but on different time periods as shown in table 3 are Moyale, Uasin Gishu, Nyando, Turkana and west pokot..

The secondary clusters reported are those that are significant at 5% level of significance and did not overlap with a reported cluster (the most likely cluster)

4.5 Tuberculosis Determinants

TB data for the year 2002 was used to analyse the effect of the covariates on Tuberculosis. A Bayesian approach with conditional autoregressive model was fitted to test the relationship of TB with the selected covariates. Before interpreting the results we looked at the stability between the posterior mean and the posterior median and found that the median was more stable than the mean. Therefore our inference is on the median and not the mean. According to the generated posterior density of β for the year 2002, all the covariate's coefficients were considered positive since their posterior median were greater than zero and their 95% credible interval did not include zero as shown in Table 4 below. This means that all these covariates were considered to have a statistically significant influence on TB. For instance the results suggest that there is an increased risk of TB in areas with high prevalence of HIV at 5% level of significance. Also increased risks of TB were observed in areas with high proportion of poverty, increased mean house hold size, unemployment, urban areas, high percentage of population that use firewood, high percentage of illiteracy, high altitude and those regions where a big percentage of the population were more than 5 kilometres away from the nearest health facility. The density plots and the trace plots in appendix 2 to appendix 11 shows the positive influence of the covariates on TB and that convergence had taken place before making any inference on the results. Convergence is considered to have taken place when most of the MCMC simulations are concentrated around the mean or median.

Table 3-Results of space time scan statistics from 2002 to 2009

District	Time frame	Observed cases	Expected cases	Relative risk	Log likelihood ratio	P-value
Most likely Nairobi	2003-2006	77190	32414.37	2.52	23467.8	<0.001
Secondary Suba , Bondo, Homabay, Migori, Siaya, Rachuonyo and Kisumu	2005-2008	63844	35693.72	1.85	9472.7	<0.001
Secondary Moyale	2004-2007	4426	787.52	5.64	4010.3	<0.001
Secondary Uasin Gishu	2004-2007	12668	9354.08	1.36	534.4	<0.001
Secondary Nyando	2003-2006	13180	10911.92	1.21	224	<0.001
Secondary Turkana	2006-2009	8701	7285.69	1.20	130.5	<0.001
Secondary West pokot	2004-2005	25.7	2238.11	1.12	15.6	<0.001

Table 4-Results of Bayesian approach using CAR specification on TB data for the year 2002.

Variable	Band width	95% credible interval	
		min	max
HIV	0.01022	0.200	3.157
Poverty	0.01065	0.200	3.145
Altitude	0.01044	0.197	3.153
Illiteracy rate	0.01034	0.201	3.156
Mean hold house size	0.01017	0.200	3.142
Employment opportunity much worse	0.01046	0.196	3.160
Employment opportunity worse	0.0151	0.199	3.151
Urban	0.009864	0.199	3.149
Distance to the nearest health facility	0.01034	0.199	3.162
Use of firewood	0.01077	0.199	3.16

5.0 DISCUSSION

In this study we have shown the spatial distribution of TB in Kenya in different districts at different time periods starting from the year 2002 to 2009. In addition we have identified areas with high occurrence of TB using Bayesian smoothed maps, spatial autocorrelation test and space time scan statistics. We have also modelled the covariates that have an influence on the occurrence of TB.

For exploratory spatial data analysis, Bayesian smoothed maps of standardized morbidity ratio were generated. This allowed visualization of spatial TB patterns in 69 districts in Kenya. We noticed that high rates of this disease were observed in the lake region and the northern part of the country with a few areas around the central region such as Nairobi, Embu and Meru south. It is also evident that some areas like Marsabit which had high rates in the years 2002 continued to decline over the years and by 2009, the rates were below the mean country rate. Districts that have consistently reported high relative risk of TB include: Moyale, Suba, Rachuonyo, Kisumu, Bondo, Migori and Nairobi among others. Nairobi could be a high risk area because it is a highly populated urban area with a number of slums leading to overcrowding and this predisposes people to TB. In addition most of the urban areas have recorded high rates of this disease over the years.

From the Bayesian smoothed maps, we were able to identify areas that need thorough follow-up if TB is to be eliminated as projected by stop TB partnership. These include districts like Moyale, Bondo, Nairobi, Turkana, Mombasa, Kisumu, Suba, Rachuonyo, and Isiolo among others. In addition, regions with emerging disease outbreaks such as Thika, Tharaka and Malindi were detected. Furthermore, districts like Turkana and Kitui that have had fluctuating TB rates over the years were also revealed.

Global Moran's I for spatial autocorrelation computed showed that districts closer to each other had similar relative risk of TB as compared to those further away. This relationship was not significant until 2007, 2008 and 2009. The local Moran's I identified spatial cluster and spatial outlier. This is very important in epidemiology since the kind of relationship in terms of TB relative risks that neighbouring districts had was portrayed. We were also able to detect areas of decreasing or increasing trends of TB in relationship to their neighbours.

Space time scan statistics are also very crucial in epidemiology as they reveal disease aggregation, the presence of statistically significant clusters and also give some information on etiologic factors. In the space time analysis computed, hotspot districts were identified for specific time periods, Nairobi being the most likely cluster significance at 5% level of significance. The results compliment those revealed in smoothed maps and cluster maps though they occurred at different time periods. In addition, the result complements those of Tiwari et al (2006), Nunes (2007), Ozunuka et al (2007) and Randremanana et al (2009) that TB occurs in clusters.

Bayesian approach, a spatial modelling technique caters for random variation as a result of differing population and model covariates while borrowing information from neighbouring region. The results have revealed that it has a high statistical power of detecting variation in disease risk. The belief that poverty is a determinant of TB has been proven in this study. These results are in line with those reported by Waaler et al (2002) and Crampin et al (2004). On the other hand HIV has a positive impact on TB in that those regions of high HIV prevalence also recorded high risk of TB. The result is similar to what was reported by Crampin et al (2004) and Lienhardt et al (2005). In addition, districts around the lake region in Nyanza province that have had high HIV prevalence are the same with increased relative risks of TB. This is also in line with the fact that TB smear negative rate has increased as a result of HIV surpassing the smear positive TB, MOPHS (2009).The mean house hold size was used as an indicator of overcrowding and the results have proved that districts with higher mean household size have a higher risk of TB keeping in mind that there are other areas not necessarily in the house that can lead to overcrowding thus exposing one to TB disease.

Altitude as a determinant of TB was found to have a positive effect. This is the opposite of what Mansoer et al (1999) and Tanrikulu et al (2008) found. This could be attributed to the fact that we included more factors than just altitude and also spatial effects were considered while modelling for this covariate. It is therefore important that this covariate is investigated further as a determinant of TB. Finally, it is also important to note that as much as SMR is usually affected by the population size, this was corrected by applying Bayesian technique through borrowing information from the neighbours.

One of the limitations of this study is that data from the districts that were created recently had to be incorporated into their previous districts so that mapping could be done. There is a possibility that some data were not captured appropriately since the data used is usually collected on a routine bases as TB patients are put on treatment. Some of the issues that could arise in the data are incorrectness, inconsistency and incompleteness. Secondly, it would have been of interest to relate the effect of gender and sex on TB in different districts in Kenya, but this was not done. Further studies could incorporate this into their model.

In conclusion, this study has revealed the spatial distribution of TB in Kenya over the years from 2002 to 2009. It has also revealed the spatial patterns of TB in different districts in comparison to their nearest neighbouring districts. Emerging districts with increased risk of TB have also been identified in addition to those districts that have consistently recorded high relative risk of TB. This information can be used by the DTLTB for planning purposes, allocation of resources' and even dissemination of TB information. It can also be used to strengthen other strategies such as poverty eradication in Kenya.

APPENDICES

Appendix 1

Mapping Tuberculosis

Accessing the district shapefile

```
library(maptools)
dist=readShapePoly("C:/Users/DWEKESA/Desktop/margaret
ndubi/ke_district_boundaries.shp")
getinfo.shape("C:/Users/DWEKESA/Desktop/margaret
ndubi/ke_district_boundaries.shp")
proj4string(dist)=("+proj=longlat +ellps=WGS84")
plot(dist,axes=TRUE)
```

Importing the TB data and merging it with the shapefile

```
data=read.csv("C:/Users/DWEKESA/Desktop/margaret ndubi/data.csv")
data
dist@data=data.frame(dist@data,data)
library(spdep)
library(MASS)
library(nlme)
library(deldir)
coords=coordinates(dist)
IDs <- row.names(as(dist, "data.frame"))
library(tripack)
dist4_nb <- tri2nb(coords, row.names = IDs)
dist5_nb <- graph2nb(soi.graph(dist4_nb, coords), row.names = IDs)
dist6_nb <- graph2nb(gabrielneigh(coords), row.names = IDs)
dist7_nb <- graph2nb(relativeneigh(coords), row.names = IDs)
```

The Poisson-Gamma Model for 2002

```
library(R2WinBUGS)

sink("bayesfit.bug")# This stores the model given by cat(" ") in the
file bayesfit.bug
cat("
model
{
# The Likelihood
for(i in 1:N)
{
observed[i]~dpois(mu[i])
mu[i]<-theta[i]*expected[i]
theta[i]~dgamma(nu, alpha)
}
# The prior
nu~dgamma(.01, .01)
alpha~dgamma(.01, .01)
}
}
",fill=TRUE)
sink()
```

The data

```
N <- length(dist$ALL.2002)
observed = dist$ALL.2002
expected = dist$E2002

data <- list("N","observed","expected")
# The data is stored in a list object that shall be called by bugs
shortly
```

MCMC initialization

```
inits=function(){list(nu=1,alpha=1)}
```

The MCMC settings

```
nc<-1
ni<-20000
nb<-10000
nt<-100
```

Start the Gibbs-Sampler, run the model in WinBUGS and save the results in the object

```
bayesfit.sim= bugs(data,inits,
model.file="C:/Users/DWEKESA/Desktop/margaret ndubi/bayesfit.bug",
parameters=c("theta","nu","alpha"),n.chains = 1,n.iter=200000,
n.burnin=10000,n.thin=100,
bugs.directory="C:/Program Files/WinBUGS14/",codaPkg=FALSE)
print(bayesfit.sim)
bayesfit.sim$summary
```

Convergence diagnostics

```
nu2002 = bayesfit.sim$sims.array[,1,1]
alpha2002= bayesfit.sim$sims.array[,1,2]

par(mfrow=c(1,3))
ts.plot(nu2002,xlab="iteration",ylab="",main="nu",col="2")
plot(density(nu2002),main="nu",col="4")
acf(nu2002,main="nu")

par(mfrow=c(1,3))
ts.plot(alpha2002,xlab="iteration",ylab="",main="alpha",col="2")
plot(density(alpha2002),main="alpha",col="4")
acf(alpha2002,main="alpha")
```

credible interval

```
dist@data$FBPGmean2002 <- bayesfit.sim$mean$theta
dist@data$FBPGmedian2002 <- bayesfit.sim$median$theta

PGLB <- data.frame(bayesfit.sim$summary[1:69,3])
PGUB <- data.frame(bayesfit.sim$summary[1:69,7])

par(mfrow=c(1,1))
```

```
PGresults <-
as.data.frame(cbind(dist@data$FBPGmean2002,dist@data$FBPGmedian2002 ,
PGLB[,1], PGUB[,1]),
row.names = as.character(dist$DISTNAME))
PGresults
plot(1,1, type="n", xlim=c(1,68), ylim=c(0.0,4),
main= "95% Cred. Int. for Bayes Poisson-Gamma Model
of Median Relative Risk",
xlab="District", ylab="Full Bayes P-G RR (median & 95% CI)", xaxt="n")
abline(h=1, lty=2)
for(i in 1:69)
{
if(PGresults$V3[i]>1 )
{
col<-gray(.4)
lty <-2
col <- "red"
lwd <-3
text(i, PGresults$V4[i]+0.23, dist$DISTNAME[i],
srt=90, col="red", cex=0.7)
}
else
{
col<-"black"
lty <-1
lwd <-2
text(i, PGresults$V4[i]+0.23, dist$DISTNAME[i],
srt=90, col=gray(.4), cex=0.7)
}
lines(c(i,i), c(PGresults$V3[i],PGresults$V4[i]),
col=col, lty=lty)
points(i, PGresults$dist$DISTNAME$FBPGmedian2009[i], pch=18, col=col)
}

library(RColorBrewer)
library(class)
library(classInt)
library(e1071)

par(mfrow=c(1,2))
plotvar=round(dist@data$FBPGmean2002,3)# The variable we are plotting
nclr=4 # The number of colors
plotclr=brewer.pal(nclr,"BuGn") # The specification of colors
class=classIntervals(plotvar,nclr,styles="quantile")
colcode=findColours(class,plotclr)
plot(dist,axes=T,xlim=c(33,45))
plot(dist,col=colcode,add=T)
plotclr=plotclr[nclr:1]
title(main="2002 PGmean for TB in Kenya")
legend(40,-2.2, legend=names(attr(colcode, "table")),fill=attr(colcode,
"palette"),cex=0.6,bty="n")


plotvar=round(dist@data$FBPGmedian2002,3)# The variable we are plotting
nclr=4 # The number of colors
plotclr=brewer.pal(nclr,"BuGn") # The specification of colors
class=classIntervals(plotvar,nclr,styles="quantile")
colcode=findColours(class,plotclr)
```

```
plot(dist,axes=T,xlim=c(33,45))
plot(dist,col=colcode,add=T)
plotclr=plotclr[nclr:1]
title(main="2002 PGmedian for TB in Kenya")
legend(40,-2.2, legend=names(attr(colcode, "table")),fill=attr(colcode,
"palette"),cex=0.6,bty="n")
```

Testing for homogeneity

```
library(DCluster)
fit <- achisq.test(dist@data$ALL.2002 ~ offset(log(dist@data$E2002)),
as(dist, "data.frame"), "multinom", 999)
fit
```

Spatial modelling with CAR specification

```
library(spdep)
library(RColorBrewer)
library(DCluster)
library(rgdal)
library(epitools)
library(R2WinBUGS)
library(coda)
```

The Model with CAR specification

```
sink("bayesfit.bug")# This stores the model given by cat(" ") in the
file bayesfit.bug
cat("
model
{
# The likelihood
for(i in 1:N)
{
observed[i] ~ dpois(mu[i])
log(theta[i])<-log(expected[i])+b0+b1*POV[i]+ u[i] + v[i]
mu[i] <- expected[i]*theta[i]
u[i] ~ dnorm(0, precu)
SMRhat[i] <- 100 * mu[i]/expected[i]
SMRraw[i] <- 100 *observed[i]/expected[i]
}
# The prior distributions
v[1:N] ~ car.normal(adj[], weights[], num[], precv)
b0 ~ dflat()
b1 ~ dnorm(0,1.0E-5)
precu ~ dgamma(0.001, 0.001)
precv ~ dgamma(0.1, 0.1)
sigmau<-1/precu
sigmav<-1/precv
}
",fill=TRUE)
sink()
```

The data

```
N <- length(dist$ALL.2002)
observed = dist$ALL.2002
```

```
expected = dist$E2002
POV=dist$PRO.POOR
hiv=dist$HIV.PREV.05
g5kms=dist$X5KMS
hhsize=dist$MEAN.H.H.SIZE
alt=dist$ALT.1
illiteracy=dist$I.L.R
firewood=dist$Firewood
urban=dist$urban
employ1=dist$E.O..M.worse
employ2=dist$E.O.worse

dist2=read.gal("dist1.gal")####dat2.gal is a rook continuity file.
dist.nb <- nb2WB(dist2)

adj = dist.nb$adj; weights = dist.nb$weights;
num = dist.nb$num
dwoutcov <- list("N","observed","expected","adj","weights","num")

data1 <- list("N","observed","expected","adj","weights","num","POV")
```

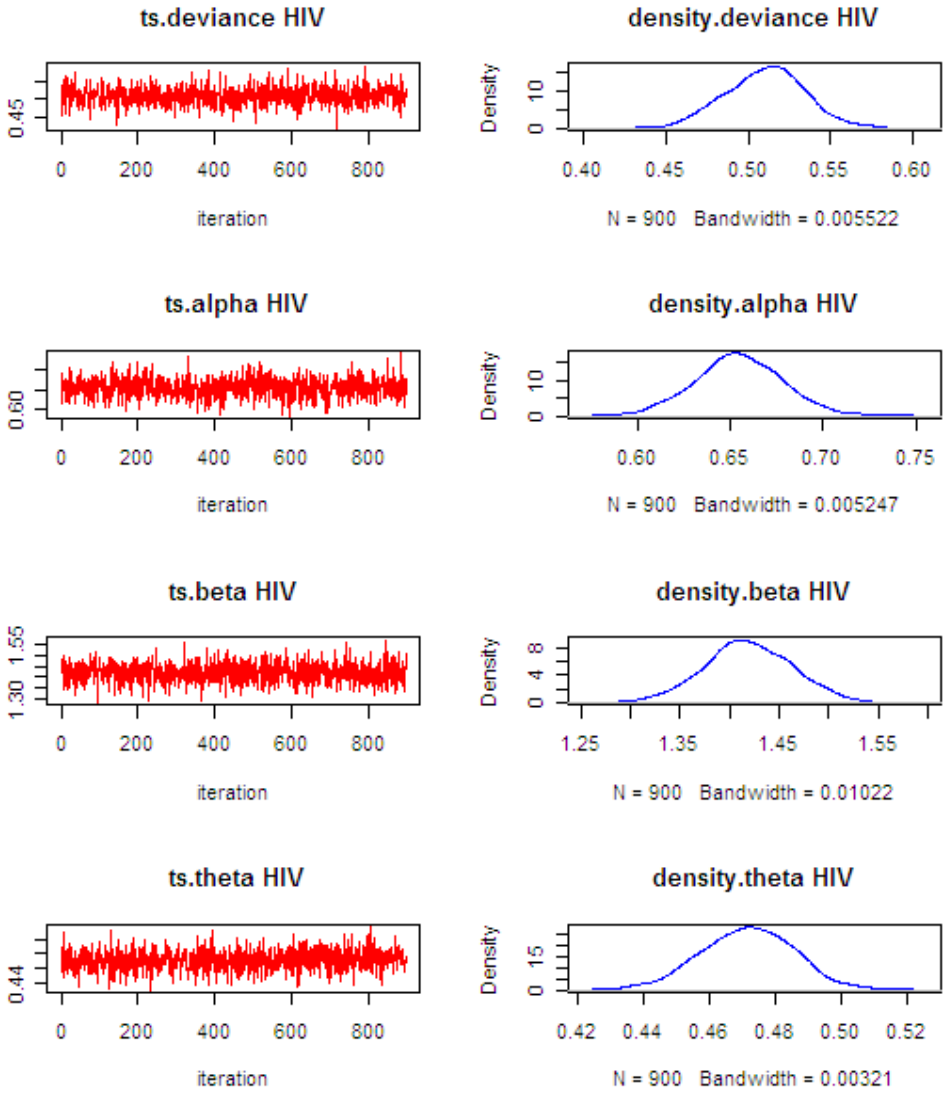
MCMC initialization

```
inits <- list(u = rep(0, N), v = rep(0, N), b0 = 0,
b1=0,precu = 0.001, precv = 0.001)


bayesfit.sim= bugs(data1,inits=list(inits),
model.file="D:/projects/margaret ndubi/bayesfit.bug",
parameters=c("theta","b0","b1","u","v","sigmau","sigmav"),n.chains = 1,
n.iter=100000,
n.burnin=10000,n.thin=100,
bugs.directory="C:/Program Files/WinBUGS14/",codaPkg=FALSE,debug=TRUE)
bayesfit.sim
```

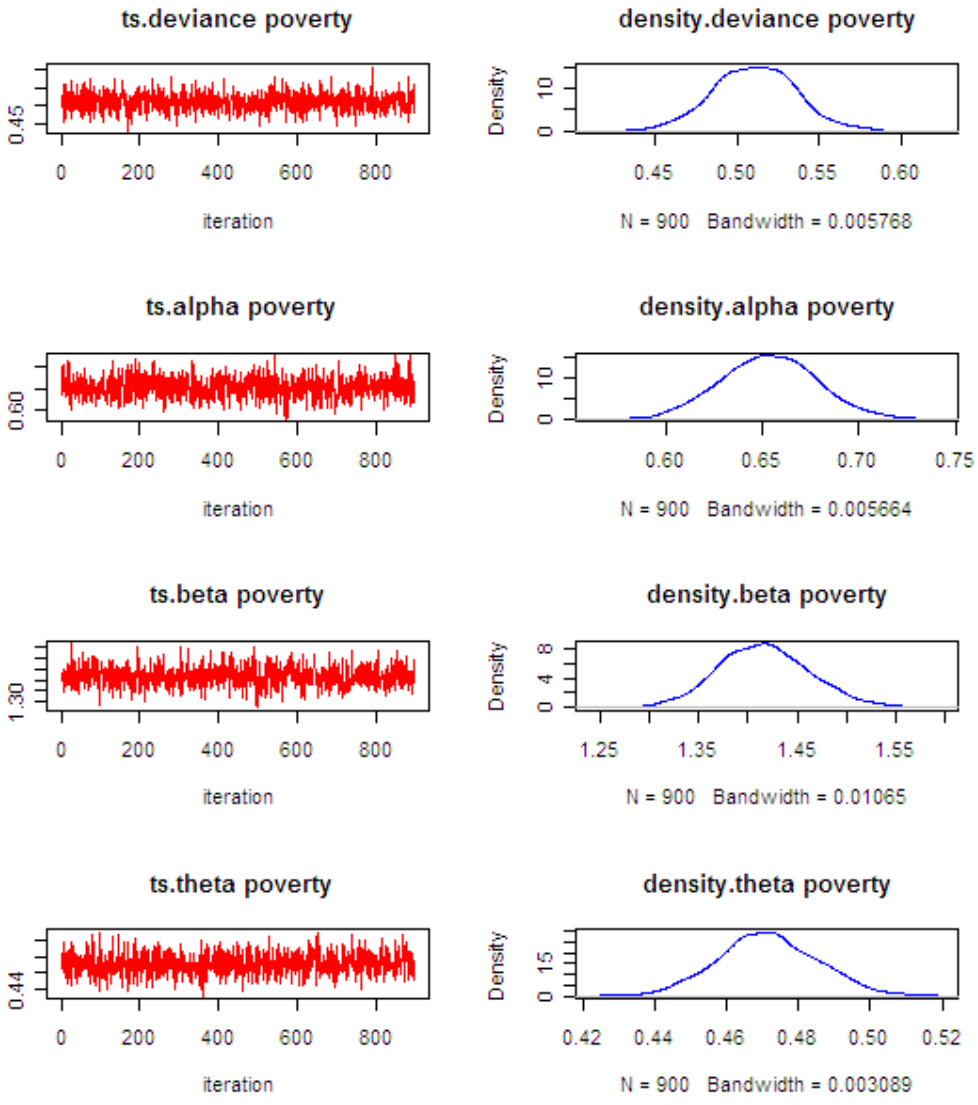
Appendix 2

Trace plots and density plots of HIV prevalence in different districts in Kenya.



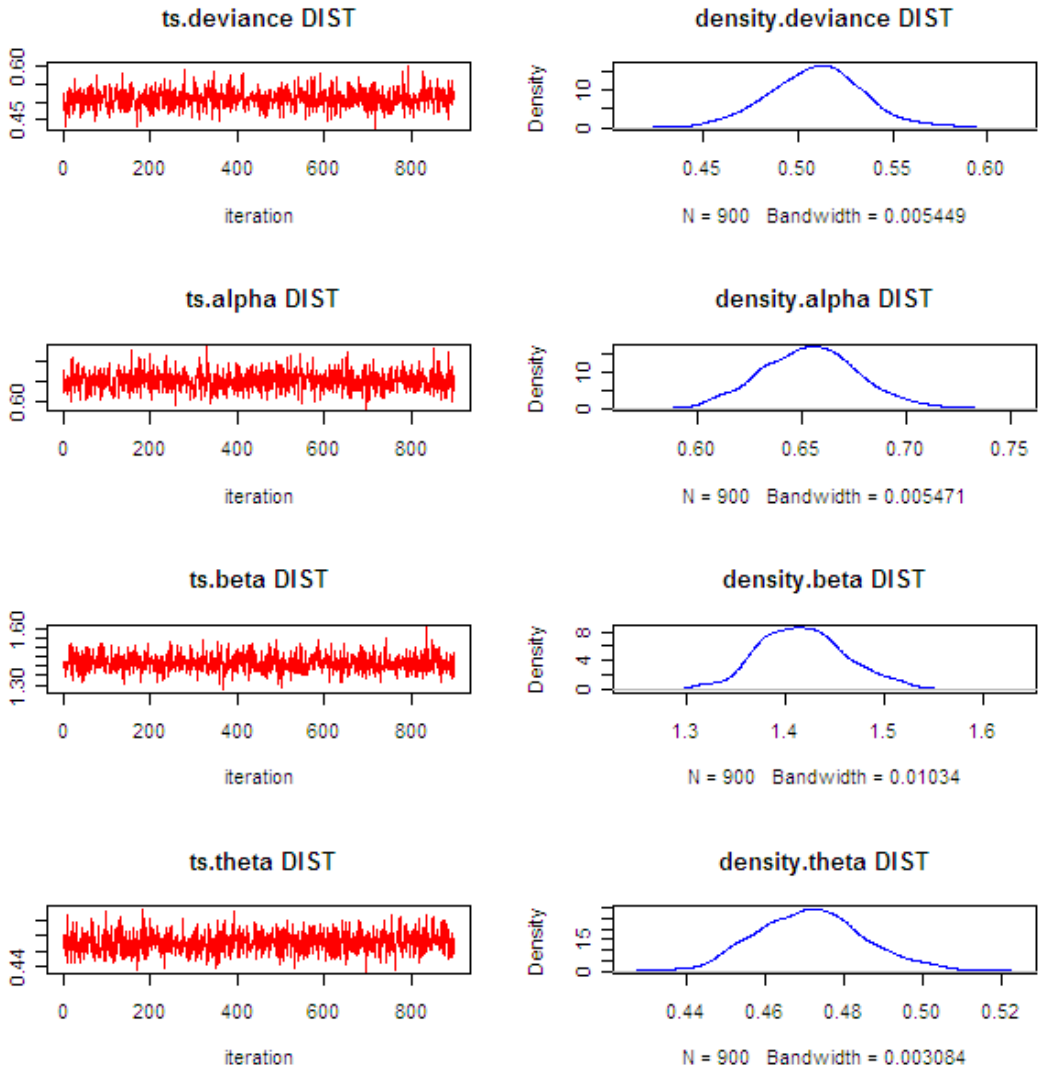
Appendix 3

Trace plots and density plots of proportion of poverty in different districts in Kenya.



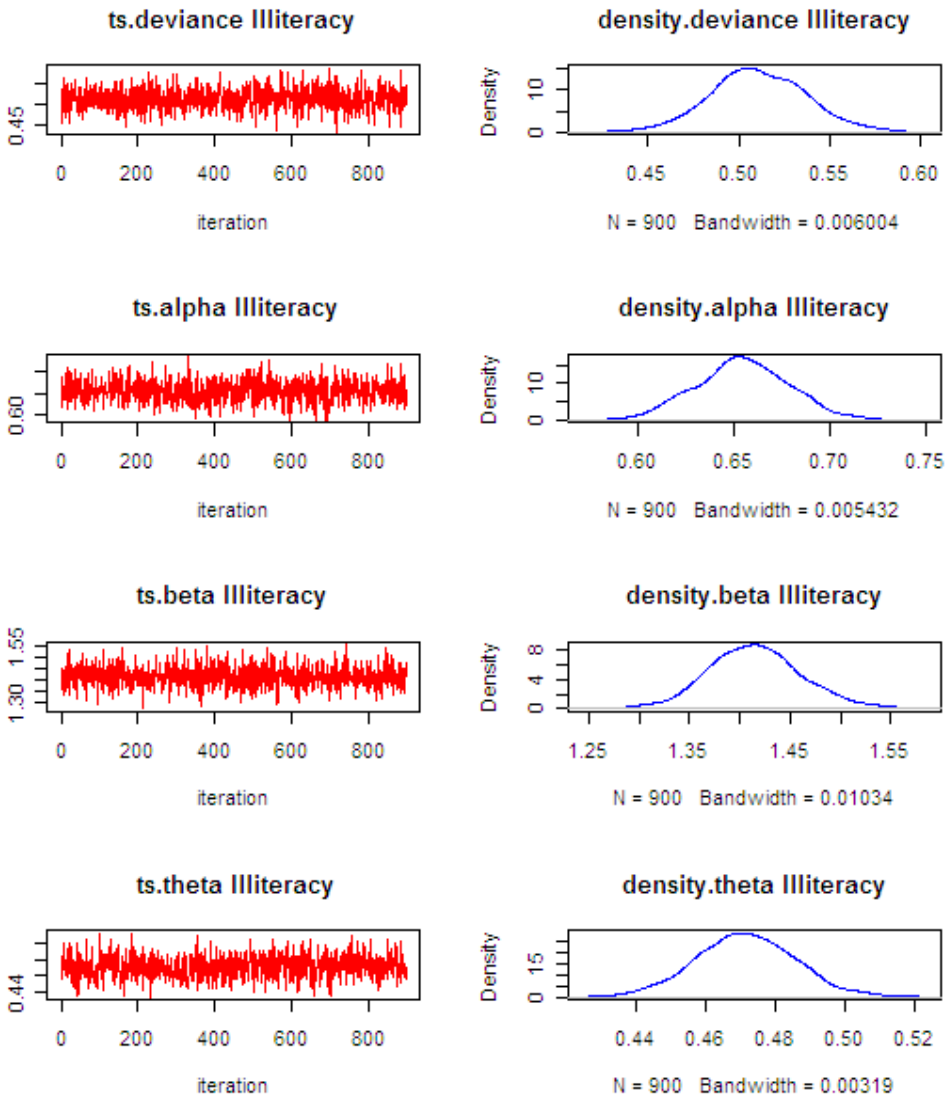
Appendix 4

Trace plots and density plots of percentage of the population that are greater than 5 kilometres away from the nearest health facility



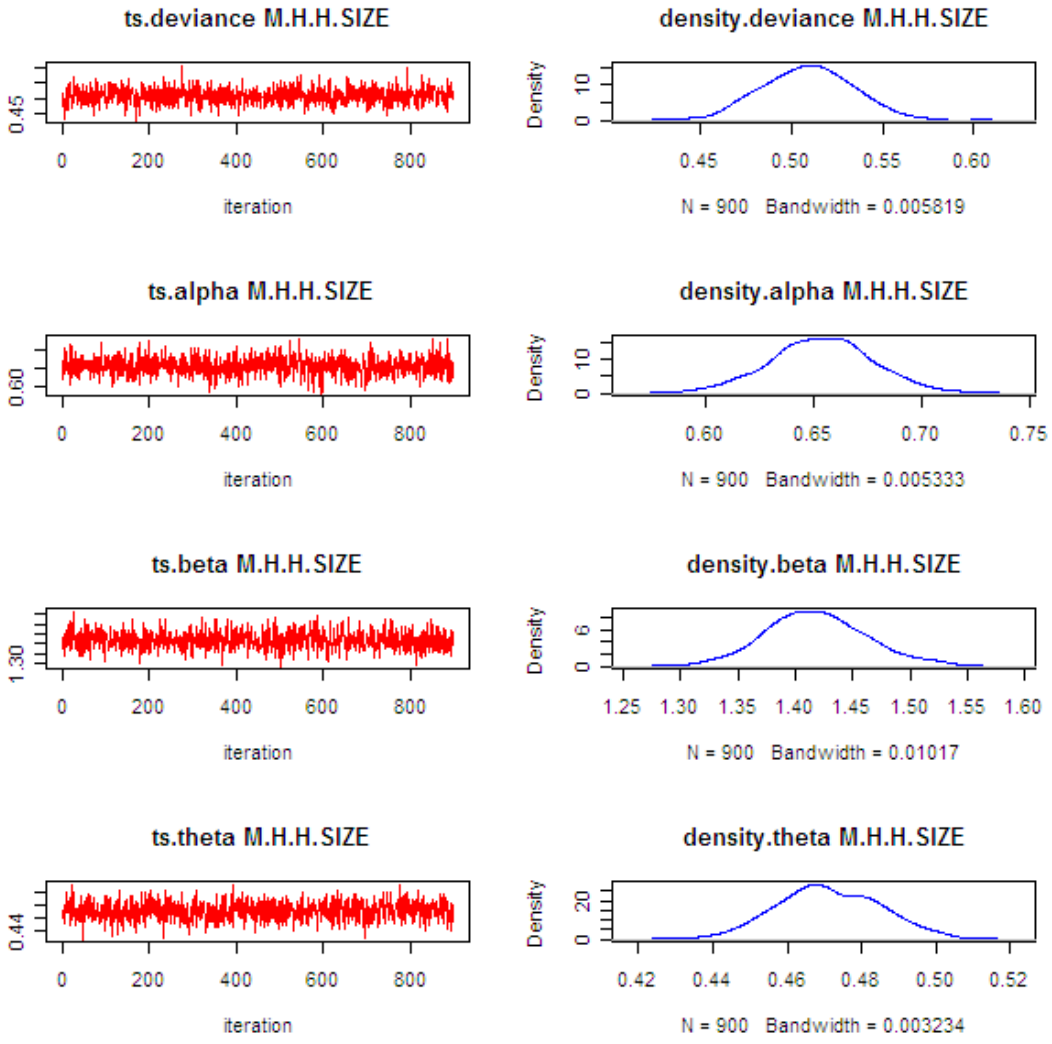
Appendix 5

Trace plots and density plots of illiteracy rate.



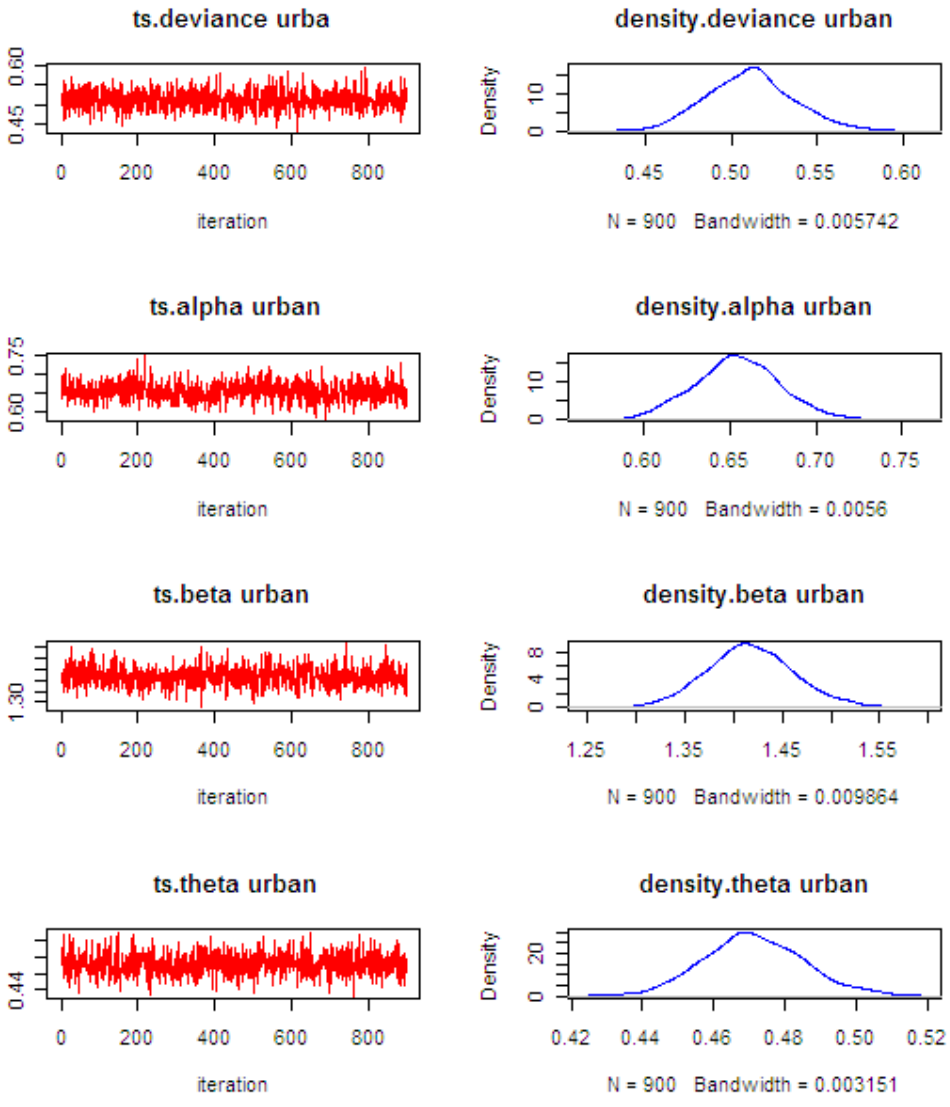
Appendix 6

Trace plots and density plots of mean household size in different districts in Kenya.



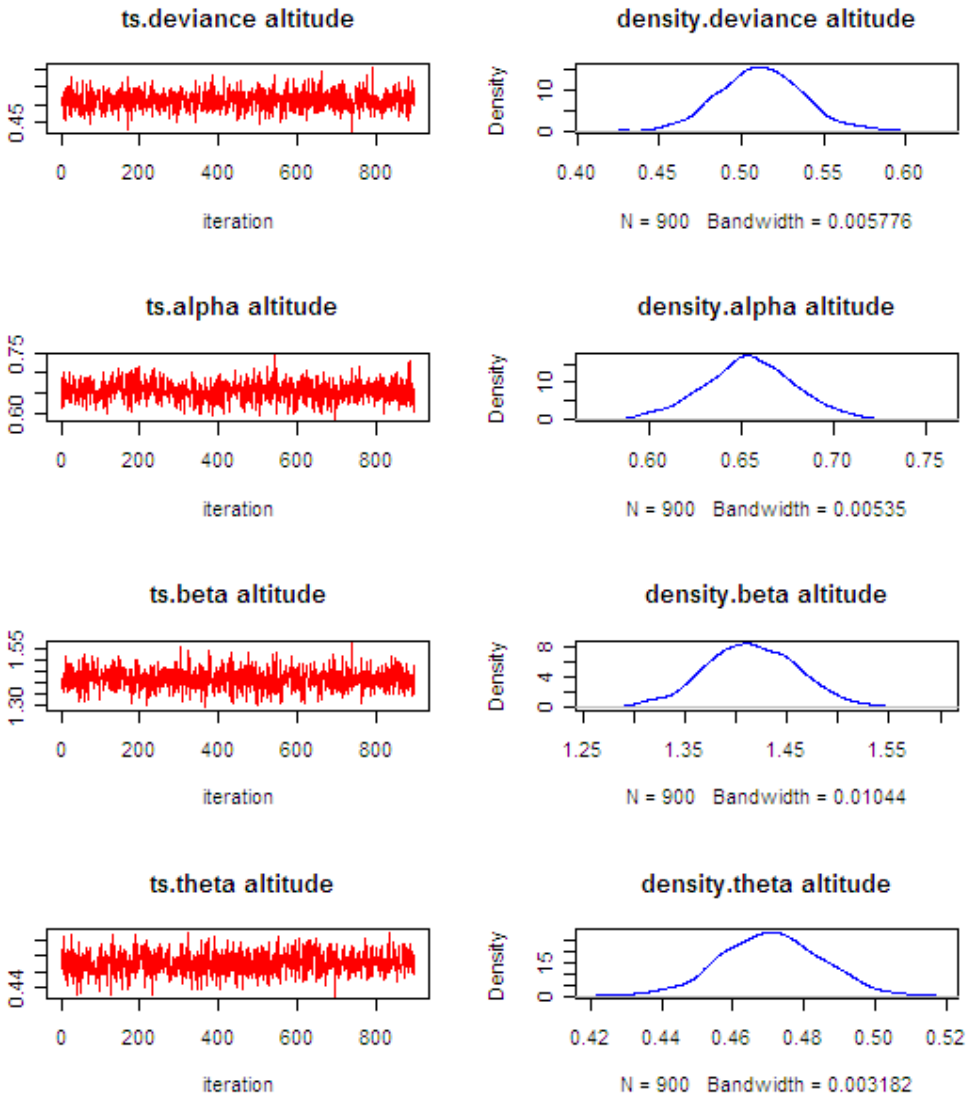
Appendix 7

Trace plots and density plots of urbanization as a risk factor for TB.



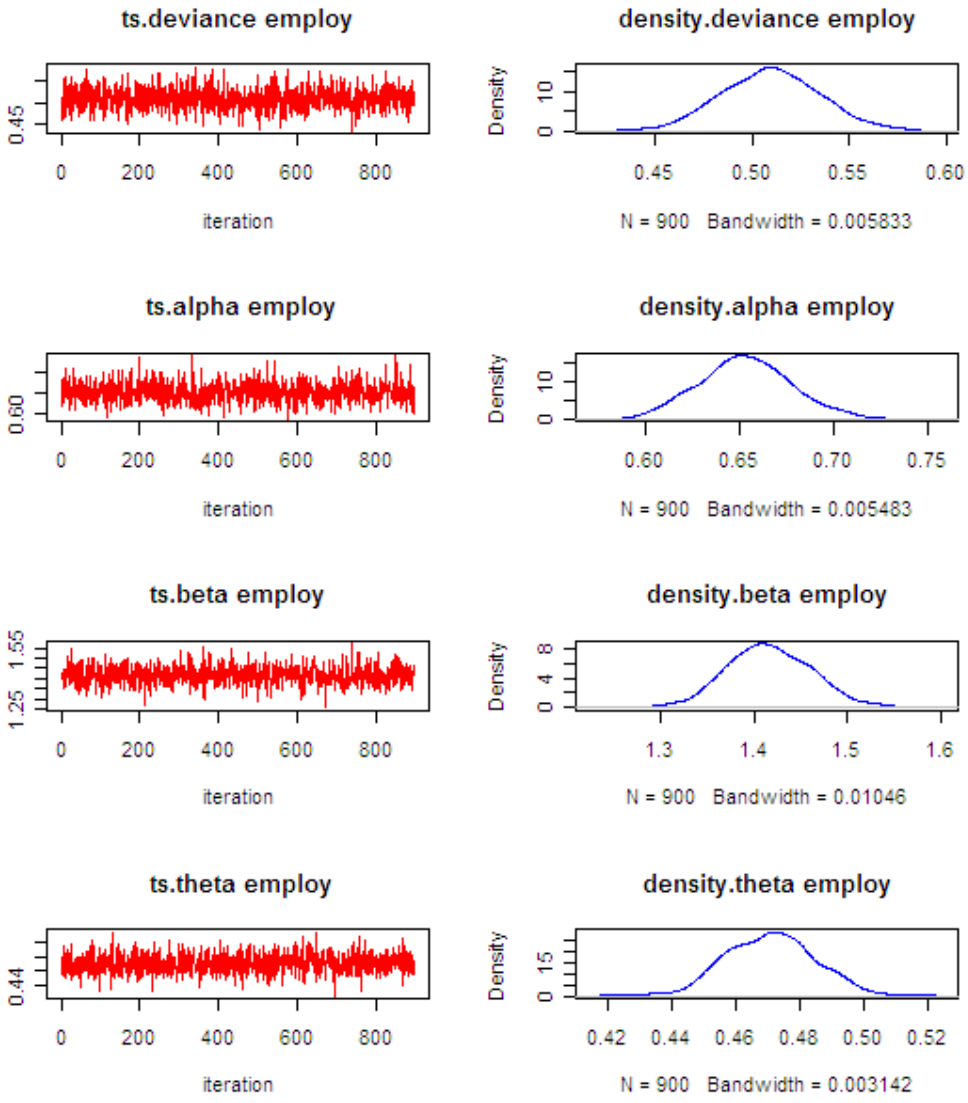
Appendix 8

Trace plots and density plots of altitude.



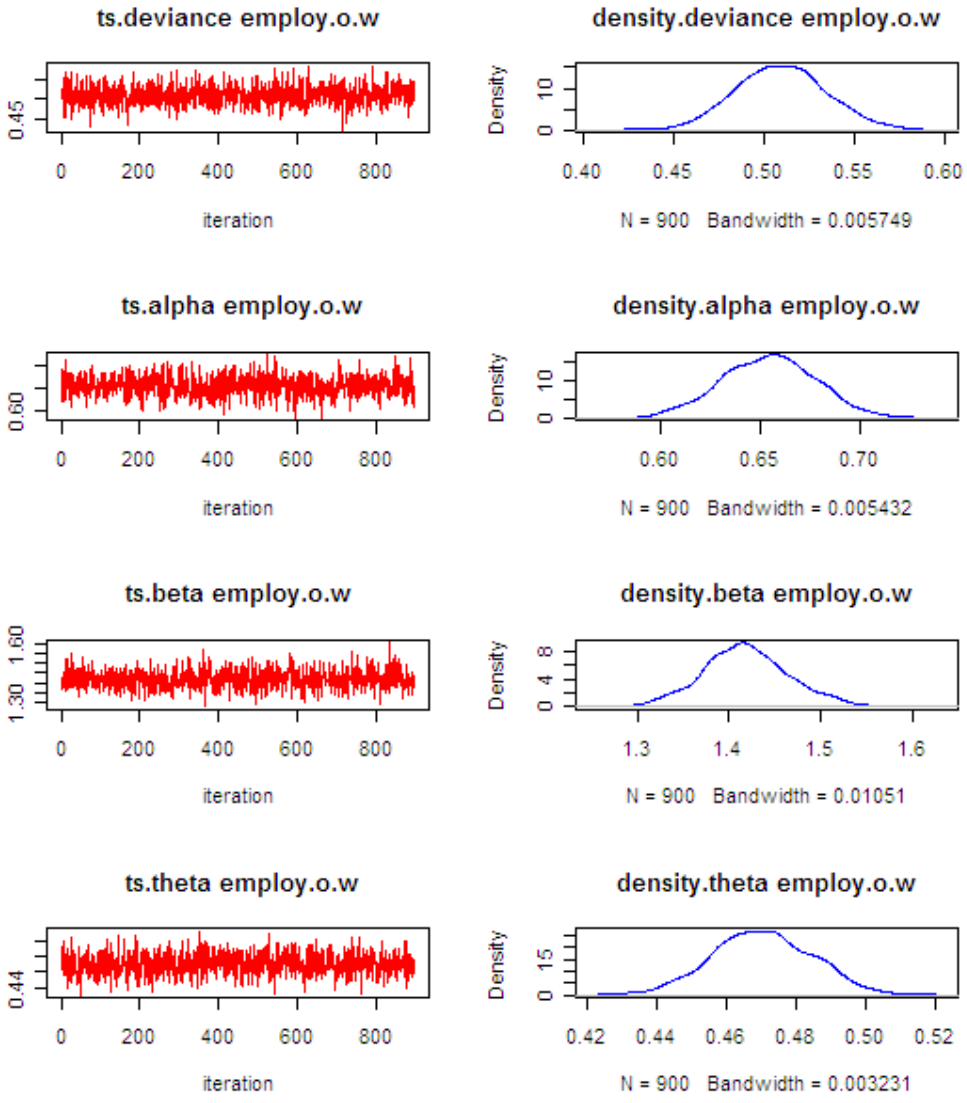
Appendix 9

Trace plots and density plots of areas with much worse employment opportunity.



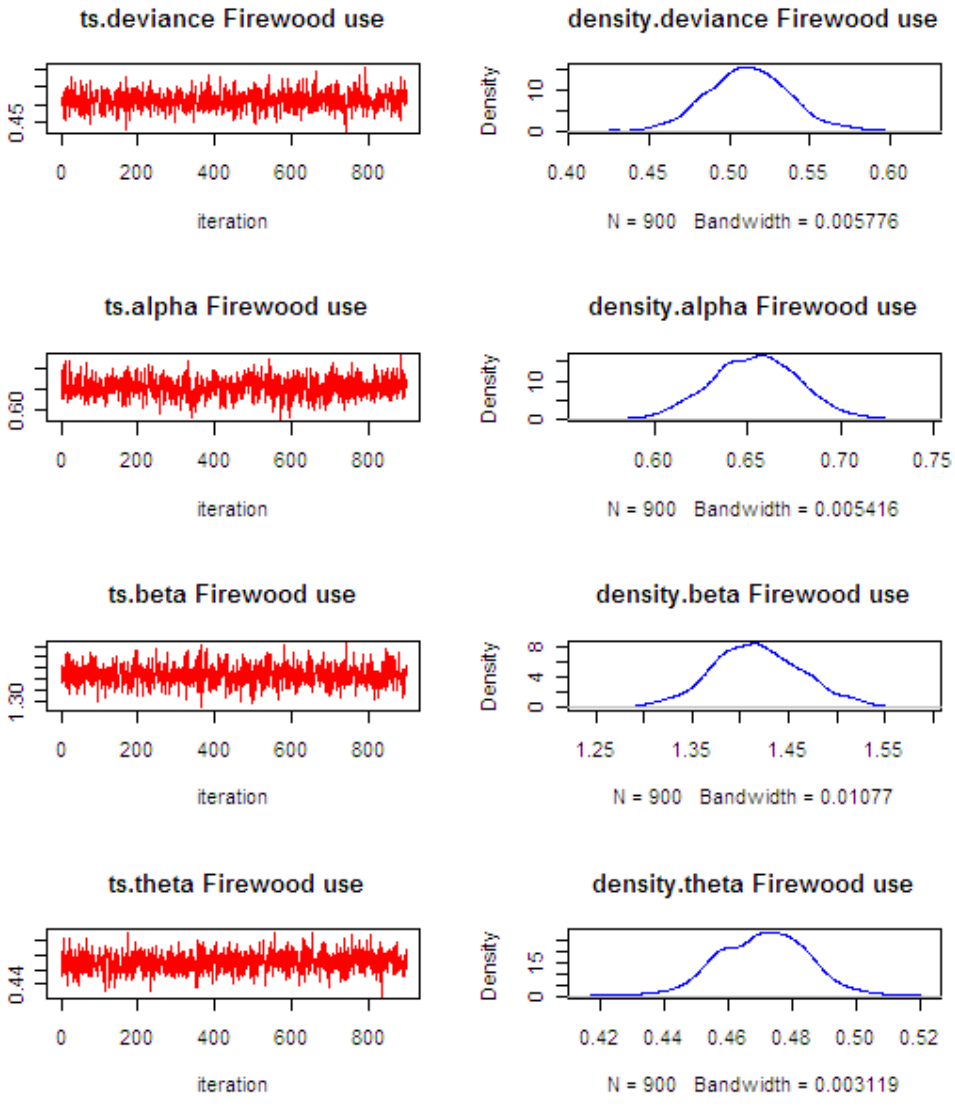
Appendix 10

Trace plots and density plots of areas with worse employment opportunity.



Appendix 11

Trace plots and density plots of proportion of the population that use firewood.



Bibliography

Assun¸c˜ao R.M, Reis.E.A. “A new proposal to adjust Moran's I for population Density.” *Statistics in Medicine* (John Wiley & Sons, Ltd.) 18 (1999): 2147-2162.

Baker M, Das D, Venugopal K, Howden-Chapman P. “Tuberculosis associated with household crowding in a developed country.” *Journal of Epidemiology and Community Health* 62 (2008): 715-721.

Barr R.G, Diez-Roux A.V, Knirsch A.C and l Pablos-Méndez A. “Neighborhood Poverty and the Resurgence of Tuberculosis in New York City, 1984–1992.” *American Journal of Public Health* 91, no. 9 (2001): 1487–1493.

Bivand R.S, Pebesma E.J and Gómez-Rubio J. *Applied spatial data analysis with R.* Springer Science and Business Media, LLC, 2008.

Callaghahan M, Cormican M, Prendegast M,Pelly H,Cloughley R, Hanhoe B and O'Donovan D. “Temporal and spatial distribution of human criptosporidiosis in the west ireland.” *International Journal of Health Geographics* 8, no. 64 (2009).

Chiang C Y, Slama K and Enarson D A. “Association between tobacco and tuberculosis.” *International journal of Tuberculosis and Lung Disease*, 2007: 258-262.

Clayton D, Kaldor J. “Empirical Bayes estimates of Age standardized Relative risks for use in disease mapping.” *Biometrics* 43 (1987): 671-681.

Coker R, McKee M, Atun R, Dimitrova B, Dodonova E, Kuznetsov S, Drobniewski F. “Risk factors for pulmonary tuberculosis in Russia: a case-control study.” *BMJ*, 2006: 85-87.

Crampin A.C, Glynn.R J, Floyd S, Malema S.S, Mwinuka V.K, Ngwira B.M.M, Mwaungulu F.D, Warndorff D.K, Fine P.E.M. “Tuberculosis and gender: exploring the patterns in a case control study in Malawi.” *International Journal of Tuberculosis and Lung Diseases* 8, no. 2 (2004): 194-203.

Davies P.D.O, Yew W.W, Ganguly D: “Smoking and tuberculosis: the epidemiological association and immunopathogenesis.” 100 (2006): 291-298.

Dye C, Lönnroth K, Jaramillo E,Williamsa BG & Raviglionea M. “Trends in Tuberculosis incidence and their determinants in 134 Countries.” 87 (2009): 683 - 691.

Fang L, Yan L, Liang S, Sake J de Vlas, Feng D. “Spatial analysis of haemorrhagic fever with renal syndrome in China.” *BMC Infectious Diseases* 6, no. 77 (2006).

Ghasemian R, Najafi N, Yadegarinia D, Alian S. “Association between cigarette smoking and pulmonary tuberculosis in men: A case-control study in Mazandaran, Iran.” *Iranian Journal of Clinical Infectious Diseases* 4, no. 3 (2009): 135-141.

Hargreaves J.R, Boccia D, Evans C.A, Adato M, Petticrew M, and John D. H. Porter. “The Social Determinants of Tuberculosis: From Evidence to Action.” *American Journal of Public Health* 101, no. 4 (2011): 654-662.

Hill P.C, Jackson-Sillah D, Donkor S.A, Otu J,Adegbola R.A and Lienhardt C. “Risk factors for pulmonary tuberculosis: a clinic-based case control study in The Gambia.” *BMC Public Health* 6, no. 156 (2006).

Jia Z, Jia X, Liu Y, Dye C, Feng C, Chen, Zhang C, Li W, Cao X,Liu W, He-Liang. “spatial analysis of Tuberculosis Cases in Migrants and Permanent Residents, Beijing, 2002-2006.” *Emerging Infectious Diseases* 14, no. 9 (2008): 1413-1419.

Kazembe.L.N. “Spatial modelling and risk factors of malaria incidence in northern Malawi.” *Acta Tropica* 102 (2007): 126-137.

Kulldorff, M. “SaTScan User Guide for version 9.0.” *[http://www.satscan.org].*, 2010.

Lienhardt C, Fielding K, Sillah J.S, Bah B, Gustafson P, Warndorff D, Palayew D, Lisse I,S Donkor, Diallo S, Manneh M, Adegbola R, AabyP , Bah-Sow O, S Bennett and McAdam K. “Investigation of the risk factors for tuberculosis: a case–control study in three countries in West Africa.” *International Journal of Epidemiology* 34 (2005): 914-923.

Lin H.O, Ezzati M, Murray M. “Tobacco Smoke, Indoor Air Pollution and Tuberculosis: A systematic review and meta-analysis.” *PLoS* 4, no. 1 (2007).

Maiti.T. “Hierarchical Bayes estimation of mortality rates for disease mapping.” *Journal of statistical planning and inference* 69 (1998): 339-348.

Mansoer.J. R, Kibuga D. K. and Borgdorff M. W. “Altitude: a determinant of Tuberculosis in Kenya.” *International Journal of Tuberculosis and Lung Disease* 3, no. 2 (1999): 156-161.

Munch Z, Lil Van S.W.P, Booysen C.N,Zietman H.L, Enarson D.A, Beyer N. “Tuberculosis transmission pattern in a high a high incidence area: a spatial analysis.” *International Journal of Tuberculosis and Lung Diseases* 7, no. 3 (2003): 271-277.

Nunes.C. “Tuberculosis incidence in Portugal: Spatio temporal clusterong.” *International Journal of Health Geographic* 6, no. 30 (2007).

Onozuka .D, Hagihara .A. “Geographic prediction of tuberculosis in Fukuoka , Japan, Using space time scan statistics.” *BMC Infectious Diseases* 7, no. 26 (2007).

Randremanana R.V, Richard V, RakotomananaF , Sabatier P, Bicout D.J. “Bayesian mapping of pulmonary tuberculosis in Antananarivo, Madagascar.” *BMC Infectious Diseases* 10, no. 21 (2010).

Randremanana R.V, Sabatier P, Rakotomanana F, Randriamanantena A and Richard V. “Spatial clustering of pulmonary tuberculosis and impact of the care factors in Antananarivo City.” *Tropical Medicine and International Health* 14, no. 4 (2009): 429-437.

Sitienei J, Wenyenga H and Kipruto H. *Annual Report.* Division of Leprocy, Tuberculosis and Lung Disease, Ministry of Public Health and Sanitation,The Republic of Kenya, 2009.

Souza W.V, Ximenes R, Maria F. M. Albuquerque, Lapa.M T, PortugaJ.P l, Maria L. C. Lima and Celina M. T. Martelli. “The use of socioeconomic factors in mapping tuberculosis risk areas in a city of northeastern Brazil.” *American Journa of Public Health* 8, no. 6 (2000).

Tanrikulu A.C, Acemoglu.h ,Palanci Y and Dagli E.C. “Tuberculosis in Turkey: High altitude and other socio-economic risk factors.” *Royal Institute of Public Health* 122, no. 6 (2008): 613-619.

Tiwari N, Adhikar.M.S C, TewariA and Kandpal V. “Investigation of geo-spatial hotspots for the occurrence of tuberculosis in Almora district, India, using GIS and spatial scan statistic.” *International Journal of Health Geographics* 5, no. 33 (2006).

Touray K, Adetifa I.M, Jallow A, Rigby J,Jeffries D, Cheung Y.B, Donkor S, Adegbola A and Hill P.C. “Spatial analysis of tuberculosis in an Urban West African setting: is there evidence of clustering?” *Tropical Medicine and International Health* 15, no. 6 (2010): 664-672.

Uthman A.O, Yahaya I,Ashfaq K and Uthman M.B. “A trend analysis and sub-regional distribution in number of people living with HIV and dying with TB in Africa, 1991 to 2006.” *International Journal of Health Geographics 2009* 8, no. 65 (2009).

Waaler, H.T. “Tuberculosis and poverty.” *International Journal of Tuberculosis and Lung Disease* 6, no. 9 (2002): 745-746.

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

“World Health Organization, WHO Report 2010. Global Tuberculosis Control: WHO/HTM/TB/2009.411.”

“World Health Organization:WHO Report 2010Global Tuberculosis Control: WHO/HTM/TB/2010.7.”

“World Health Organization; WHO Report 2008, Global and Tuberculosis Control ; Surveillance, Planning and Financing:WHO/HTM/TB/2008/393.”

Zaman K, Yunus M, Arifeen S, Baqui A.H, Sacks D.A, Hossain S, Rahim Z, Ali M, Banu , Islam M.A, Begum N, Begum V, Breiman R.F and Black R.E. “Prevalence of sputum smear-positive tuberculosis in a rural area in Bangladesh.” *Epidemiological Infection* 134 (2006): 1052-1059.