Bayesian Disease Mapping in The Presence of Under-reporting

MS300-0001/2015,

Oti-Boateng Emmanuel,

MSc. Mathematical Statistics.

A Thesis submitted to Pan African University Institute of Basic Science, Technology and Innovation in partial fulfillment of the requirement for the award of the degree of Master of Science in Mathematics (Mathematical Statistics) of the Pan African University.

2016.

Declaration.

I do hereby declare that this is my original work and has not been presented in partial fulfillment of any other degree award in any other university.

Signature: ..... Date .....

Name:Oti-Boateng Emmanuel.

This research thesis has been submitted for examination with our approval as University Supervisor.

Signature ..... Date .....

Dr. Ngesa Owino Oscar.

Taita Taveta University

Taita-Taveta-Kenya.

(Supervisor).

Signature ..... Date .....

Dr. Osei Badu Frank.

6University of Energy and Natural Resources.

Sunyani-Ghana.

(Supervisor).

Dedication

I dedicate this work to my uncles and their wive, Mr. and Mrs. Ebenzer Appiah and Mr. and Mrs. Isaac Kwakye.

Acknowledgment

I will like to show appreciation to God for the gift of life and good health. I will also like to show appreciation to my supervisors, Dr. Oscar Owino Ngesa and Dr. Frank Badu Osei on whose shoulders I stood.

To Prof. Peter Mwita and the entire PAU staff I say a big thank you for the impact you have made in my life.

Table of Contents

List of Tables

List of Figures

Nomenclature

α Intercept Parameter

$\in$ A member of

λ Relative risk of Contracting Diabetes

$\mathbb{R}^n$ A set of Real numbers of Dimension n

β Parameters

σ Varaince Parameter

τ Variance Parameter

$N(i)$ St of Neighbours of Region i

u_{1i} Correlated Spatial Effect of the Count Data

u_{2i} Correlated Spatial Effect of Under-reporting

v_{1i} Uncorrelated Spatial Effect of Count Data

v_{2i} Uncorrelated Spatial Part of Under-reporting

CAR Conditional Auto Regressive

CI Credible Interval

DIC Deviance Information Criterion

GCM Gaussian Component Mixture

GIS Geographical Information Systems

GLM Generalized Linear Model

ICAR Intrinsic Conditional Auto-Regressive

IDF International Diabetes Foundation

IDFAFR International Diabetes Foundation For Africa

MCMC Markov Chain Monte Carlo

MLE Maximum Likelihood Estimation

Abstract

In real life situations, the values of the response variable, which is the count data is mostly under-reported. In this work, we develop a model to cater for under-reporting in count data. In particular, we allow under-reporting to vary spatially by regions and it is captured by a binomial probability. Poisson distribution is used in modeling the count response under the assumption that over-dispersion does not exist. In the case of under-reporting, it was made to also vary spatially from one unit to the other through a probability captured by a binomial distribution.

The spatial variations of the disease were divided into correlated and uncorrelated parts. When a Poisson Regression analysis was used, both the correlated and uncorrelated parts were all found to share a significant relationship with the relative risk for each region with more contribution coming from the uncorrelated part. The model obtained was applied to diabetes data in Ghana. Disease maps for the diseases are also developed for Ghana at administrative (district) level. These maps are critical and informative to policy makers. These maps allow them to target policies and use the already meagre resources well.

Chapter 1

Introduction

1.1 Background of study

Spatial disease mapping is a technique used to display the distribution and prevalence of some named disease in a given geographical area. This has long been a part of public health, epidemiology and the study of disease in human populations (Koch, 2005). This allows epidemiologists to better understand the interaction between humans and the environment. Many scientists have utilized this scientific approach in proposing realized solution of some of the world's epidemiological problems; one of such people is Dr. John Snow who, in 1854, identified the residence locations of cholera deaths during the London epidemic and connected a certain public water pump to the accumulation of cholera cases which eventually led to the closing of the pump in question. Before this, epidemiologists were only interested in mapping the locations of disease cases and rate until the advent of improved scientific technologies like Geographic Information Systems (GISs) (Moore et al., 1999).

Disease mapping has seen a tremendous metamorphosis after the invention of the modern computing and (GISs) (Moore et al., 1999). GIS allows the capture, manipulation, analysis and the graphical display of all kinds of spatial or geographically referenced

data transcending beyond the traditional statistical method of scatter plot which involves the displaying of the degree of relationship between multiple respondent and response variables (Moore et al., 1999). Another key element in spatial data captured by the introduction of new technology is clustering or disease clustering.

Diseases tend to cluster because the movement of humans within a given locality can not be likened to a random case. Clustering come in two forms, either local (the interest lies more in the characteristics of the clusters which comes in the form of size, location and intensity) or global (where clustering is studied in relation to disease in the whole geographical area under study) (Tango, 2010). Based on the idea of clustering, spatial patterns are identified in the disease risk of the population (Wartenberg, 1999). This was also confirmed by (Almani et al., 2008), who reported that changes in etiological factors (environmental variables) have direct impact on diseases. Studying clustering in data gives an upper hand to epidemiologists because there is some relationship between spatial pattern and the demographic and environmental variables (Besag and Newell, 1991a). Although there are so many methods of testing for global clustering, Moran Index is the most widely used (Moran, 1950). Scientists based on this idea of clustering to identify traces of spatial patterns induced by some of these variables. However, there are other significant factors such as the nature of the population at risk. Again, transmission processes induce spatial patterns in data, especially for infectious diseases where the mode of transmission is by contact. For the case of non-infectious diseases (cardiovascular and diabetes cases), spatial variation in disease outcomes is induced by spatial variation in the demographic structure leading to higher rates in areas with individuals at high risk (Moore et al., 1999). Contextual variables such as poverty indices and administrative approaches also contribute to disease occurrence in regions.

Most importantly, the nature of data collection impacts on the spatial effect underlying the observed spatial patterns of the disease (Moore et al., 1999). Normally, i.e. in

real situations, there is under-reporting of cases on disease due to variability of case reporting at the local level thereby creating some filter in the true underlying pattern of the disease. Places with efficient reporting programs appear up in the observed program; this is not the case of most places in the world and it worsens as one travels down Africa (Moore et al., 1999). Under-reporting in the public health may be as a result of fear of stigmatization, inadequate funds to seek medical expertise, inadequate knowledge on the disease, lack of confidence in the existing health institution, failure of successive medication, to mention but a few. These anomalies have devastating effects, some of which are, it produces biased estimations for count models (Ye and Lord, 2011).

In cases where under-reporting occurs, it is difficult to estimate the true state of some diseases based on the reported cases. Estimation is more complicated and complex when the factor responsible for under-reporting is immeasurable. For instance, when it comes to drinking habbits of respondents, men are more likely to respond than women. In this example, incidence rate is likely to be biased if the gender variable is deliberately or indeliberately ignored.

In order to account for the above problems, we combined a well developed model-based approach and an efficient method of estimation called the Bayesian method of estimation. Bayesian method will be the most efficient especially in our case where data on covariates are not available. The unavailability of influencing covariates will create an escape window where the missing covariates will be treated as latents. The model devloped will be validated using diabetes data from Ghana.

Diabetes is a major public health issues in terms of both morbidity and mortality. Diabetes is currently at the epidemic level with 70 percent of those infected living in low and middle income economies (WHO, 2014). About 87 million people have diabetes in the world and more than 22 million people in the Africa Region; by 2035 this figure will almost double (IDF, 2014). Prevalence of diabetes in Africa as at 2007

was nearing the 10.4 million mark (WHO, 2014). Ghana is one of the 32 countries of the IDFAFR region. Ghana has its fair share of diabetes mellitus at monumental score of 450,000 cases of diabetes in 2014 and cost per person with diabetes stands at 148.8USD (IDF, 2014).

Several studies have been done on the subject, however, a nationwide comprehensive work is yet to be done considering cases of diabetes. Most importantly, a statistical model is yet to be developed in the field of disease modeling with under-reporting captured by binomial probability and varying spatially for all units under consideration. This adds to the uniqueness of the work. The purpose of this study is to develop a spatial model with significant consideration of under-reporting for diabetes cases in Ghana. The model will be used in plotting disease map using available diabetes data.

1.2 Statement of problem

Any process which operates in space creates patterns (Ripley, 1977). This means that virtually all human activities have the tendency to create patterns. However, these patterns can not be seen or observed with the human eye thereby necessitating the need for a special scientific method called spatial pattern analysis. Spatial pattern analysis is aimed at identifying, understanding and describing the process of spatial patterns. Many models have been drawn to model this property in count data. However, a dominant flaw in such models is as a result of the assumption that the values of response variable is correctly reported which is not the case in real life situations. Most of the times we are faced with the case of under-reporting of count data which negatively affects the correct modeling of real data. Under-reporting comes about when reported cases are less than the true state of cases in the given geographical area. This phenomenon subjects policy makers to difficulty when investing in areas affected. In this work, an interest in the modeling of count data lagged with under-reporting is

considered. We assume that under-reported cases vary spatially through a probability captured by the binomial distribution and also under the assumption that each individual event is reported. Also, a more reliable method of estimation will be employed other than the usual Maximum Likelihood Estimator method. In this case, the identified parameters will be identified and then estimated using the Bayesian method of estimation. Bayesian method has an advantage over the frequentist due to the inclusion of prior knowledge aside information from data. Real data on diabetes cases is used to complement the prior distribution in computing the posterior distributions. A Poisson distribution is then used in the over all modeling of the count data. The approach will be compared to existing ones so as to conclude on the more efficient one.

1.3 Objectives

1.3.1 Main objective

To develop a spatial model for disease counts in the presence of under-reporting using Bayesian estimation.

1.3.2 Specific Objectives

1. Develop a spatial model to cater for varied under-reporting in the spatial units (districts).
2. To apply the model above to estimate relative risk of diabetes in Ghana (using data from the districts) by a method of Bayesian estimation.
3. Develop disease maps for diabetes cases for all districts in Ghana.

1.4 Significance of the study

Africa is lagging behind in infrastructure and basic social amenities in the face of scarce resources. There is therefore the need for policy makers to make meaningful and pragmatic decisions in the way of spending tax payers' money. In Ghana, government is most times at a loss as to which particular area of health needs more and urgent attention, especially when most cases of non-infectious diseases are not reported instead traditional solutions are sought, which brings about the problem of under-reporting. Disease mapping can be used to answer such pressing questions and also identify indicators that directly or indirectly fuel disease transmission. Specifically, this study will acknowledge a better method of estimating the parameters associated with spatial analysis by the application of Baye's theory, opening new doors for research in the area.

1.5 The Scope of the study

The aim of this study is to develop a spatial model for count data lagged with under-reporting, determine and map the relative risk of contracting diabetes in Ghana by a method termed disease mapping. Data describing diabetes cases were retrieved from the Ghana Health Service, for all districts of Ghana. Current issue of the data,i.e. all entries for data for 2015 was not available except for former years. Also, it was not easy to acquire these data in covariates. Using this available data, the developed models will be validated and compared to other existing models. Also, disease map will be produced for all districts in the country.

1.6 Outline of the thesis

The thesis contains five chapters. The first chapter which introduces the subject matter like spatial statistics, disease mapping and also scope of diabetes cases in Africa.

Chapter two talks about the some literature on the field of spatial statistics, Bayesian method, Generalized Linear Models and disease mapping. In chapter three, a theory that justifies the choice of Poisson distribution over others is proved. This is then extended into the methodology where the Besag, York and Mollie model was extended and further on, a Bayesian method of estimation was used in the parameter estimation. Chapter four begins with the explanation of of the data used. Also, results are displayed in tables and disease maps. The thesis is then ended with the last chapter which is mainly on conclusion and recommendation for further research works.

Chapter 2

Literature Review

2.1 Introduction

The degree of inconvenience as a result of diabetes has attracted the attention of not only policy makers, but also peasant farmers in the remotest of communities. Quite a number of interventions have been used in addressing these problems which include but not limited to the health sector response, priority interventions for prevention (Exercising), treatment and care in the health sector, operationalizing the priority interventions – strengthening health systems, investing in strategic information etc. To boost the effectiveness of these interventions, there is the need to locate areas where these diseases thrive and persist the most, so that policy makers can know where to invest resources. One of the methods used in this identification is called spatial disease mapping, which aims to identify the risk of contracting some named diseases in a given geographical area. There has been quite a number of application of this procedure (disease mapping); notable among them are (Zayeri et al., 2011) who presented the geographical map of malaria by identifying some of the important environmental factors of the disease in Sistan and Baluchistan province, Iran. In their paper, the registered malaria data was used in the computation of the Standard Incidence Rates (SIR). A

geographical mapping of malaria incidence rates were mapped with subsequent environmental factors. In their work, except for rainfall humidity, elevation, average minimum temperature and average maximum temperature had a postivie relationship with malaria SIRs in this entire province. Also, (Moraga and Lawson, 2012) used an alternative model i.e. the Gaussian Component Mixture (GCM) model instead of the proper or improper Conditional Autoregressive (CAR) in disease mapping. In their paper, a review of CAR and GCM models are investigated in the univariate sense. Also, an addition of spatial effects as random effects are investigated. The method of estimation was the Bayesian method. Although the above literature elaborates on efficient estimation and disease mapping methods, one dominant mistake is the assumption that observed data is a true reflection of the reality. Gibbons et al., 2014 investigated and identified the advantages of using efficient, reliable surveillance and notification systems associated vital factors for monitoring public health and disease outbreaks. The mose recent work in this area is (Neubauer et al., 2016) who confirmed that in the presence of under-reporting, parameter estimation fails. They corrected this anomaly by extending the binomial model i.e. the use of mixied models in modeling the under-reported phenomenon although they employed a frequentist method of parameter estimation which is disadvantaged when compared to the Bayesian method. In this part of the world, especially in developing countries, not all disease cases are reported. This brings up a gap in the number of cases called under-reporting in count data making it difficult for policy makers to tackle the problem. Also, developing countries are faced with inadequate resources in the collection and colation of health data subkecting data analysis to the mercy of latent variables. These among other reasons necessitated a more efficient method of estimating the disease incidence in a given geographical area. One of the recent works on under-reporting is Gamado et al., 2014 who worked on modeling under-reporting in epidemics by considering the stochastic Markov SIR epidemic in which various reporting processes are incorporated. In their

work, they were able to show that, excluding under-reporting breeds a case of under estimating the infectious rate. They also incorporated under-reporting and developed suitable models by allowing a reporting probability which depended on time. Bayesian method of estimation was then used owing to the fact that data reported was incomplete.

In this paper, we reconsider the work of Gamado et al. (2014). Especially allowing under-reporting to vary through a probability, except that under-reporting was made to vary from one region to the other through a probability captured by a binomial distribution. This is in addition to the fact that the incidence rate in the Poisson distribution is also allowed to vary spatially making this work unique.

2.2 Generalized Linear Model (GLM)

One very important aspect of epidemiology is being able to separate determining variables (independent variables or covariates) from the response variable (dependent variables) in a regression analysis. Normally, a regression model is used in drawing this relationship when the error term follows a normal distribution. Otherwise, a more flexible or extended case of regression model called Generalized Linear Model is used. The above-mentioned methods only explore the first order effects of the contributing covariates on the mean of the disease outcome leaving out small-scale but very significant variations that could be as a result of interactions between boundary-sharing-units, i.e. spatial auto-correlation. The assumption normally used is that, any spatial traces lined in the data is accounted for by the spatial properties in the covariates. It suffices to say that, in the case whereby a particular covariate, varying spatially, deliberately or indeliberately omitted, then there will be bias in the estimation of the covariates in question (Draper and Smith, 1998). The cumulative effect of this will be a spatial auto-correlation in the residual process, thereby underestimating the standard error of the covariates resulting in overestimating of the statistical significance of the test (Cressie,

1993).

In our case, it is without ambiguity that our random variable or indepenent variables are the spatial effects and the response variable becomes the number of cases in diabetes. Our main objective is involves the estimation of under-reporting in our count data. In effect, a robust model that accommodates this phenomenon is highly recommended. Also changes in our respondents X does not trigger linear changes in our response, Y, as such we introduce a $link\,function$ (arbitrary function) to correct this. In our case we employ *Poisson regression analysis*, a type of GLM, to model the count data and predict the incidence rates.

In the problem statement, the case of under-reporting was clearly identified. Winkelmann (1996) proposed a Poisson regression model where the spatial effect is captured using a binomial distribution of varying spatial probabilities for each region. A *logarithm* or a our *link function* whereby the response variable assumes a Poisson distribution is employed to connect the incidence rate and the independent variables (spatial effects). This can be written mathematically by supposing that, $\mathbf{x} \in \mathbb{R}^n$ is a vector of independent variables, then the model takes the form $\log \mathbf{E}(Y \mid \mathbf{x})) = \alpha + \beta' \mathbf{x}$ where $\alpha \in \mathbb{R}$ *and* $\beta' \in \mathbb{R}^n$ (Feller, 1968). Sometimes this is written more compactly as,

$$\mu = \mathbf{E}(Y \mid \mathbf{x}) = e^{\theta' \mathbf{x}}, \tag{2.2.1}$$

where θ is simply α combined to β', $\mathbf{x}$ is a $n+1$ dimensional vector. The Bayesian method will be used in estimating θ because, this method has an advantage of correcting confounding problems that are not incorporated in data . Here, β is taken to be a a vector of coefficients of some covariates, $\mathbf{x}$. Some of the known methods of estimating β are the Maximum Likelihood Estimation and the iterative Least Squares (Feller, 1968) given as,

$$\sum \left\{ y_i^* - \exp\left(x_i' \beta\right) \right\} x_i'. \tag{2.2.2}$$

The mean and variance in Poisson distribution may bring about the problem of over dispersion in certain cases.

2.3 Review of the Besag, York and Mollie Model

The Besag,York and Mollie were the pioneers when it comes to modeling spatial effects in count data. To model the spatial effects in count data, it is supposed that, λ_i represents the relative risk for region i with respect to a standard population, also let y_i and E_i denote the observed counts of the disease and the expected count in region i respectively. The count data can be modeled by the Poisson distribution below;

$$y_i \sim Poisson(E_i\lambda_i); \tag{2.3.1}$$

Based on the assumption that the log of relative risk of disease can be broken down into a spatially structured component u_i and a spatially unstructured component v_i which can be written mathematically as;

$$log(\lambda_i) = u_i + v_i. \tag{2.3.2}$$

Besag and Newell (1991a) noted that in most cases, one of the random effects usually dominates the other. If u_i is stronger than v_i, then the estimated risk will show spatial structure and if v_i is stronger than u_i then the consequence will be to shrink the estimated means towards the overall mean. Besag and Newell, 1991a assumed that u and v were independent with the following priors:

$$p(v \mid \tau) \propto \tau^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\tau}\sum_{i=1}^{n} v_i^2\right\} \tag{2.3.3}$$

and

$$p(u \mid k) \propto k^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i}\sum_{j\in N(i)} \left(u_i - u_j\right)^2\right\}. \tag{2.3.4}$$

$v \sim WN(0,\tau)$ and u follows Gaussian Markov Random Field (GMRF) process with variance k and n being the number of districts under study and $N_{(i)}$ is the set of neighbors of region i. Based on the regions of shared border (which sometimes taken to be

the actual distance between locations or centroids of locations), (Besag and Newell, 1991b; Ngesa et al., 2014a) modeled the conditional distributions of each u_i as;

$$p\left(u_i \mid \mathbf{u}_{-i}\right) \sim N\left(\frac{\sum_{j \in N(i)} u_j}{d_i}, d_i^{-1} k\right), \tag{2.3.5}$$

with,

$$\mathbf{E}\left(u_i \mid \mathbf{u}_{-i}\right) = \frac{\sum_{j \in N(i)} u_j}{d_i}, \tag{2.3.6}$$

and

$$Var\left(u_i \mid \mathbf{u}_{-i}\right) = \frac{k}{d_i}, \tag{2.3.7}$$

where d_i is the number of neighborhoods of region i.

This conditional distribution for u is called the Intrinsic Conditional Auto-regressive (ICAR) prior distribution (Besag and Newell, 1991a; Ngesa et al., 2014a).

2.4 Bayesian analysis.

This is the method of statistical estimation based on Bayes theorem, whereby posterior distribution about an unknown parameter is borne based on prior information and data (Walsh, 2002). Bayesian method is preferred to frequentist method due to its reliance on Markov chain Monte Carlo methods, (MCMC). This method, that is the MCMC method, used to be computationally exhausting when it was initially proposed. The MCMC method decomposes complicated estimation problems into simpler problems that rely on conditional distributions for each parameter in the model (Gelfand and Smith, 1990).

Here, the unknown (fixed) parameters are identified and some inference in the form of their distribution of domain of existence, is identified (Berger, 2013). This is then complimented by the likelihood function which is mostly computed out of data generated by some random variable (Starkweather, 2011). The likelihood function is derived by the popular Maximum Likelihood Estimation method. From these two information,

the posterior distribution is calculated using the Bayes theorem (Walsh, 2002; Berger, 2013). According to this theorem, the posterior is said to be proportional to the product of the likelihood function and the prior with the constant of proportionality being an inverse of the normalizing constant.

Another interesting way of doing this is to assume that, the uncertainty about the true parameter follows some probability distribution termed prior. Suppose θ follows a prior probability distribution function (pdf),$p(\theta)$. Now for a given θ, the pdf of X can be written as $p(x,\theta)$. The joint distribution of (θ,X) can be written as $p(\theta)p(x \mid \theta)$(Berger, 2013) . After seeing the data, the belief about θ can be updated by calculating the conditional distribution of θ given $X = x$. In that case, our posterior can be given as;

$$p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{m(x)}, \tag{2.4.1}$$

where,

$$m(x) = \int p(\theta)p(x \mid \theta)\, d\theta, \tag{2.4.2}$$

is the marginal distribution of X. By Bayes theorem, the posterior distribution is proportional to the product of the prior and the likelihood. Simulations are then generated out of which empirical distributions are derived for the true parameters then topped up with the generation of summary of the empirical simulates using basic statistics.

Confounding effects of latent parameters are eliminated leading to a better estimate of the posterior distribution of the unknown parameter unlike classical statistics. Given a specified multivariate distribution, a sequence of observations can be generated by a sampling method that employs Markov chain Monte Carlo algorithm, especially when direct sampling fails. Two of such methods are Gibbs sampling and Metropolis Hastings algorithm. A typical example of software that employs this algorithm is WinBUGS (Ntzoufras, 2011).

Some Bayesian analysts find the fact that the posterior, $p(\theta \mid X)$ inference depends on

the prior, $p(\theta)$, a great disadvantage. Different researchers might have different arguments about the uncertainty surrounding the prior and this normally leads to different conclusions. A good way of solving this is by sticking to standard priors for the unknown parameters (Berger, 2013). Normally, these prior distributions are chosen to be non informative so that the domain of existence gives room for large dispersion. Non informative priors are also chosen so that all the information about the posterior comes from the data and not the prior. Also, hyper parameters are set in such a way that, the precision is at its maximum best. Gamma distribution is normally chosen as the distribution for the precision as variance only appear as non negative numbers (Walsh, 2002; Smith and LeSage, 2004; Berger, 2013; Ngesa et al., 2014a).

It is always advised that the prior distribution (whether weak or strong) is chosen with caution suiting the situation at hand. The type of prior depends on the strength of belief we have in it. Weak prior arises when we don't have enough information on the prior and scientists reacts to this by extending the range of existence of the true parameter (Walsh (2002); Ntzoufras (2011)). This increases the level of influence of the likelihood on the posterior distribution. Otherwise, the strong prior is recommended, meaning that we have strong evidence about the property of the prior distribution and confine our domain of existence to be small. When this happens, the posterior is highly influenced by the prior(Walsh (2002); Ntzoufras (2011)). This scenario normally ends up in the cases of conjugate priors.

2.5 Gibbs sampling

Gibbs sampling or sampler is a type of Monte carlo Markov Chain (MCMC) simulation which can be used to obtain a sequence of observed values. A histogram is then plootted out of these values to arrive at a desired distribution.

The sequence of observed values can be used to approximate the joint, marginal distri-

bution of some subset variable distribution by way of a histogram and also to determine the expected values of the variables. When some of the values of variables are known, they are treated as latent parameters and hence do not need to be sampled (Walsh (2002)). Gibbs sampling, by virtue of being a Markov chain, generates samples which correlates with neighbor samples. To penalize this effect, thinning the generated chain of samples by considering specified n^{th} value. In addition, the initial generated samples of the chain (the burn-in period) may not accurately represent the desired distribution. In that case we apply the burn -in period by ignoring the first j^{th} values. Below are the steps for the algorithm.

Suppose we want to obtain k samples, $\mathbf{X} = (x_1, \ldots, x_n)$ from a joint distribution $p\,(x_{1,\cdots,}x_n)$.

Denote the i^{th} sample by $\mathbf{X}^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})$.

Let $\mathbf{X}^{(0)}$ to be our initial values.

To get the $(i+1)^{th}$, we pick each component variable, which is the $(i+1)^{th}$ value given the $\left(i^{th}\right)$ value written compctly as, $\left(x_j^{(i+1)} \mid x_{j-}^{(i+1)}\right)$. From the distribution of that variable conditioned on all other variables, making use of the most recent values and updating the variable with its new value as soon as it has been sampled. This requires updating each of the component variables in turn. If we are up to the j^{th} component we update it according to the distribution specified by $p(x_j|x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$. Note that we use the value that the $(j+1)^{th}$ component had in the i^{th} sample not the $(i+1)^{th}$.

Repeat the above step k times.

2.6 Metropolis-Hastings algorithm

This is also a type of Markov chain Monte Carlo (MCMC) method used in generating a sequence of random samples from a probability distribution from which direct

sampling is difficult. From the generated samples, histograms can be constructed representing the distributions of the unknown parameters. Below is the illustration of the algorithm.

Construct Markov Chain $Y^{(t)}$with stationary distribution $f(y)$.

At time t, generate next value $Y^{(t+1)}$ by going through the following steps;

Proposed step: Sample Y from the proposed distribution;

$Z \sim q\left(z \mid Y^{(T)}\right)$

Acceptance: With Probabilities :

$\alpha\left(Y^{(t)}, Z\right) = min\left\{1, \frac{f(z)}{f(z^t)} \frac{q\left(Y^{(t)} \mid Z\right)}{q(Z) Y^{(t)}}\right\}$

Set;

$Y^{(t+1)} = Z,$ for the acceptance. Otherwise,

$Y^{(t+1)} = Y^t,\ for$ the rejection.

2.7 Deviance information criterion (DIC)

The introduction of Markov chain Monte Carlo (MCMC) has revolutionized the way we fit models to real world data or complex data (Gilks, 2005). Our ability to fit models provides us with answers to the numerous behavior of collected data such as; whether the addition of random effect will go a long way to cater for over-dispersion. Normally, model comparison begins from defining a model of fit statistics (deviance) and complexity as a result of the number of unknown parameters in the model. Models are compared based on these two quantities of maximum likelihood since complexity is accompanied by best fit, Akaike's information criterion or either of the above mentioned (Aitkin, 1991). Other Bayesian model comparison which uses Bayes factor approximation also requires the specification of the number of parameters in each

model Spiegelhalter et al. (1998).

Deviance is one of the methods of goodness-of-fit statistics for models based on the generalization of the idea behind the sum of squares of residuals in ordinary least squares. This helps in model-fitting achieved by maximum likelihood. Many authors suggested different statistics to be used as a measure of fit but (Dempster (1997)) suggested plotting the posterior distribution with the deviance under each model. Here fit is identified as the posterior mean of the deviance and *complexity*, p_D, is the difference between the posterior mean of the deviance and the deviance based on the posterior means of the parameters. These results are easily computed by MCMC analysis. The DIC is then computed by adding the complexity to the fit which is used for the model comparison as suggested by (Spiegelhalter et al. (1998)). The best fitting model is the one with the smallest DIC value (Ngesa et al. (2014b)). In cases where the difference in DIC between the models is not above 5, we select the simplest model, that is, model with both few parameters and random effects (Spiegelhalter et al. (1998); Ngesa et al. (2014b)). WinBugs version 1.4 can easily generate DIC values and can run into negatives.

Chapter 3

Methodology

3.1 Proposed model

Count data, a term used in statistics, arises when one is faced with an observation that only comes in the form of positive integers and also as a result of counting. To model such data when observations are taken as count values, one has a choice of choosing from one of the following; Poisson, Binomial and Negative Binomial distribution. There is therefore the need to justify the choice of a particular distribution when faced with modeling count data. In this thesis, we choose the theoretical approach over the graphical representation because of its advantages. The theoretical approach presents one with a general perspective whiles the graphical approach only focuses on some given data.

We achieve this by assuming a fixed time period, t and let y_i^* represent the total number of events that occurred in unit i. Also, we assume that y_i^* conditioned on the covariates is Poisson distributed with mean given in Equation (2.2.1), $i \in \mathbb{N}$. In order to correctly model this phenomenon of under-reporting, it is imperative to note that, the reported cases, y does not represent the true state of count data in the unit. As such, y only represents a fraction of y^* and it is under-reported.

We suppose that,

$$p\left(y_i|y_i^*,\lambda_i\right) \sim Bin\left(y^*,\lambda_i\right), \tag{3.1.1}$$

which can also be writen as;

$$p\left(y_i|y_i^*,\lambda_i\right) = \left(\begin{array}{c} y^* \\ y \end{array} \right) \lambda_i^{y_i} \left(1-\lambda_i\right)^{y_i^*-y_i}. \tag{3.1.2}$$

The probability of an individual reporting an event is λ_i. It is also supported by the following assumptions; that the process of an individual reporting an event is memory-less and assumed constant. Winkelmann and Zimmermann (1993) then noted that y can be deduced in many ways. One is, the number of reported cases can be assumed to represent the true number of cases in the unit, $y_i = y_i^*$. Another way is when $y_i = y_i^* - n$ where $n < y_i^*$ to be any number of non-reported cases. This makes the marginal distribution of the number of reported cases, y_i, to be computed as;

$$p\left(Y_i = y\right) = \sum_{y^* \geq y}^{\infty} \frac{E^{y^*} e^{-E}}{y^*!} \frac{y^*!}{(y^*-y)!y!} \lambda^y \left(1-\lambda\right)^{y^*-y}, \tag{3.1.3}$$

$$p\left(Y_i = y\right) = \sum_{y^* \geq 0}^{\infty} E^{y^*} \left(1-\lambda\right)^{y^*} \frac{\left(\lambda E\right)^y e^{-E}}{y!}. \tag{3.1.4}$$

The Left Hand Side of Equation (3.1.3) can be equated to;

$$p\left(Y_i = y\right) = \left\{ \left[1 + E\left(1-\lambda\right) + \left(E\left(1-\lambda\right)\right)^2 + \left(E\left(1-\lambda\right)\right)^3 + \cdots \right] \left(\frac{e^E \left(\lambda E\right)^y}{y!} \right) \right\}, \tag{3.1.5}$$

with,

$$\left.\begin{array}{c}\left(\dfrac{1}{E(1-\lambda)-1}\right)>0,\ where\ 0<\lambda<1\\ E(1-\lambda)>1\end{array}\right\}$$

$$p(Y_i=y)=\left(\frac{1}{E(1-\lambda)-1}\right)\left(\frac{e^{-E}(\lambda E)^y}{y!}\right),. \tag{3.1.6}$$

$$p(Y_i=y)=\left(\frac{e^{-\lambda E}(\lambda E)^y}{y!}\right),\ y\geq 0. \tag{3.1.7}$$

Hence Equation (3.1.7) confirms that the observed counts follow a Poisson distribution with mean λE. Normally, an assumption of independence can be attached to the process of λ. However, in the case where this assumption is relaxed, λ can be expressed to depend on some covariates and it is treated as a random effect (Winkelmann and Zimmermann (1993)). The new model then takes a form of a Poisson model and can be treated as one with an unobserved heterogeneity. This suffices that, the mean of the Poisson distribution can be given as;

$$log\mathbf{E}(Y|X_i,\lambda_i)=X'\beta+\ln\lambda_i,\ where\ \lambda_i\in[0,1]. \tag{3.1.8}$$

In Equation (3.1.8), $\ln\lambda_i$ is strictly negative and its effect is additive making the marginal expectation of Y_i to be computed as;

$$\mathbf{E}(Y_i|x_i)=E[\mathbf{E}(\lambda_i)]=\mu E. \tag{3.1.9}$$

The Variance of Y_i in Equation (3.1.8) can be written as;

$$\begin{array}{c}Var(Y_i|x_i)=\mathbf{E}(Var(Y_i|x_i,\lambda_i))+Var(\mathbf{E}(Y_i|x_i,\lambda_i))\\ Var(Y_i|x_i)=\mathbf{E}(Y_i|x_i)+E^2Var(\lambda_i)\end{array} \tag{3.1.10}$$

$$Var(Y_i|x_i)=\mu+E^2Var(\lambda_i). \tag{3.1.11}$$

Since $Var(Y_i|x_i) > \mathbf{E}(Y_i|x_i)$, we conclude that under-reporting just like unobserved heterogeneity leads to over-dispersion. Over-dispersion means that there was a higher variation in the data than predicted.

Again, ignoring the presence of over-dispersion and with reference from Equation (3.1.7), the count data is modeled by supposing that, if $\mathbf{Y}=(y_1, y_2, \cdots, y_n)$ are the number of counts in each region of some disease in Ghana and E_i is the expected number of counts in unit i, then the following distributional assumption can be made in the Poisson model. Suppose that the under-reporting varying spatially through the probability π_i, $\left(\pi_i \neq \pi_j \text{ for } i\neq \text{j}\right), i = 1, 2, \ldots, n$, captured by the binomial distribution and under the assumption that each individual event is reported. Then the proposed spatial Poisson regression model for count data can be written as;

$$P(Y = y_i) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} exp(-\mu_i)}{y_i!}, y_i = 0, 1, 2, \ldots \tag{3.1.12}$$

$$\mu_i = \phi_i \pi_i, \tag{3.1.13}$$

$$\phi_i = E_i \lambda_i, \tag{3.1.14}$$

With reference to Equations (3.1.13 and 3.1.14), Equation (3.1.12) will transform into;

$$L(\mu; y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} \frac{(\lambda_i E_i \pi_i)^{y_i} \exp(-\lambda_i E_i \pi_i)}{y_i!}, y_i = 0, 1, 2, \ldots \tag{3.1.15}$$

$$\propto \left(\prod_{i=1}^{n} \frac{(\lambda_i \pi_i)^{y_i}}{y_i!} \right) \exp\left(-\sum_{i=1}^{n} \lambda_i E_i \pi_i \right), y_i = 0, 1, 2, \ldots \tag{3.1.16}$$

and,

$$\ell(\mu; y_1, y_2, \cdots, y_n) = \ln(\mu; y_1, y_2, \cdots, y_n) = \sum_{i=1}^{n} \left\{ y_i \ln(\lambda_i E_i \pi_i) - \sum (\lambda_i E_i \pi_i) - \ln(y_i!) \right\}, \tag{3.1.17}$$

$$log(\lambda_i E_i \pi_i) = log\lambda_i + logE_i + log\pi_i, \tag{3.1.18}$$

where,

$$log\lambda_i = u_{1i} + v_{1i} + X'\beta, \tag{3.1.19}$$

In Equation (3.1.19), u_{1i} and v_{1i} are the structured and unstructured spatial effects respectively in the count data.

$$logit\,(\pi_i) = u_{2i} + v_{2i}, \tag{3.1.20}$$

In Equation (3.1.19), u_{2i} and v_{2i} are the structured and unstructured spatial effects respectively in the under-reported cases, π_i is the probability of under-reporting in region i and it is different for each region.

Also, λ_i is the relative risk in unit i. In equation 3.1.18, the effect of the predictors on the count data is multiplicative

, not additive (Rodrıguez (2007)). The mean count of region i can be represented by and the prior distribution of the correlated parts of the spatial effects can be given as;

$$p\,(u_{1i} \mid \mathbf{u}_{-1i}) \sim N\left(\frac{\sum_{j\in N(1i)} u_{1j}}{d_{1i}}, d_{1i}^{-1}k_{1i}\right), \tag{3.1.21}$$

$$p\,(u_{2i} \mid \mathbf{u}_{-2i}) \sim N\left(\frac{\sum_{j\in N(2i)} u_{2j}}{d_{2i}}, d_{2i}^{-1}k_{2i}\right). \tag{3.1.22}$$

In Equations (3.1.21 and 3.1.22), d_{1i}, d_{2i} are the number of neighboring units, $N\,(1i)$ and $N\,(2i)$ are the set of neighbors of $1i$ and $2i$ respectively and k_{1i}, k_{2i} are unknown and they will have to be determined by choosing appropriate hyper parameters when using Bayesian method. To achieve these, the unknown parameters are assigned non-informative prior. In most cases, the Gamma distribution is used to model precision parameters (inverse of variance parameters) with known hyper parameters α_i and δ_i with expectation $\frac{\alpha_i}{\delta_i}$ and variance $\frac{\alpha_i}{\delta_i^2}$. WinBugs performs its analysis based on inverse gamma distribution for the variance (Ntzoufras (2011)). The prior distributions for the uncorrelated parts of the spatial effect can be given as;

$$v_{1i} \sim WN\left(0, \sigma_{1i}^2\right) \tag{3.1.23}$$

$$v_{2i} \sim WN\left(0, \sigma_{2i}^2\right). \tag{3.1.24}$$

The $\sigma_{1i}^2, \sigma_{2i}^2$ in Equtions (3.1.23) and (3.1.23) represent fixed variances.

It is also imperative for us to determine the log-likelihood of the coefficients of the covariates by considering μ_i to be dependent on solely $\mathbf{X}\beta$ for the reason that, the rest of the variables will be omitted when the derivatives are applied. We also omit $\ln y_i!$ for the same reason; we arrive at,

$$\ln L(\beta) = \sum_{i=1}^{n} \left\{ y_i \left(u_{1i} + v_{1i} + u_{1i} + v_{2i} + X'\beta \right) - \exp\left(u_{1i} + v_{1i} + u_{2i} + v_{2i} + X'\beta \right) \right\} = 0, \tag{3.1.25}$$

$$= \sum_{i=1}^{n} \left\{ y_i \left(u_{1i} + v_{1i} + u_{1i} + v_{2i} + X'\beta \right) - \exp\left(u_{1i} + v_{1i} + u_{1i} + v_{2i} + X'\beta \right) \right\} = 0. \tag{3.1.26}$$

When the derivatives with respect to the $\beta's$ are worked out for Equation (3.1.26), we arrive at;

$$X'y = X'\hat{\mu}. \tag{3.1.27}$$

In the equation above, X is a matrix containing rows of observations and columns of predictors. Also, y is a vector of response variables whiles $\hat{\mu}$ represents vector of estimated $\beta's$ estimated using the known MLE. This information is then combined with priors to compute the posteriors in a method called Bayesian method of parameter estimation. This is discussed extensively below.

With reference to Equations (3.1.19 and 3.1.20), the candidate models are developed, which increase in complexity with addition of models some random effects and parameters.

$$Model\,1: \log \mu_i = \log E_i + \alpha_0 + u_{1i} + v_{1i}, \tag{3.1.28}$$

$$Model\,2: \log \mu_i = \log E_i + \alpha_0 + u_{1i} + v_{1i} + u_{2i}, \tag{3.1.29}$$

$$Model\,3: \log \mu_i = \log E_i + \alpha_0 + u_{1i} + v_{1i} + v_{2i}, \tag{3.1.30}$$

$$Model\,4: \log \mu_i = \log E_i + \alpha_0 + u_{1i} + v_{1i} + v_{2i} + u_{2i}. \tag{3.1.31}$$

The models above were validated using data retrieved from the Ghana Health Service, an independent institution charged with the collection and collation of data in all aspect of health importance, at the district level. Data on diabetes was collected, summed and recorded on monthly basis by the district health offices. The estimated population for each district, for the study period was obtained from the Ghana Statistical Service. Data on this morbidity is available for all districts of Ghana. The period of consideration is 2014 as the agency did not have all data for later years. There are no missing data of any kind.

Model estimation was carried out using a Bayesian approach with every parameter being assigned prior distributions. To be precise, a non informative normal prior was assigned to the offset parameter, α_0 whiles the variance parameters are assigned inverse gamma distributions. The thesis was carried out under the assumption that covariates are not available. Win Bugs version 1.4 was used in the implementation (Spiegelhalter et al. (2003)) phase. A double chain of Markov chain Monte Carlo (MCMC) iterations of 120,000 were ran with initial of 10,000 left out as the burn-in period and then every tenth sample value considered for arriving at the convergence of the estimates the remaining 11,000 samples. The decision on convergence was arrived at based on the behaviour of our trace plots and auto-correlation plots of the MCMC output. The models were compared using the Deviance Information Criterion (DIC) as proposed by (Spiegelhalter et al. (2003)). The best fitting model is one with the smallest DIC

value.

3.2 Parameter estimation

Bayesian method was used in estimating the unknown parameters. This is the method of statistical estimation whereby posterior distribution about an unknown parameter is borne based on prior information and data. Suppose a random variable $Y = (y_1, y_2, \cdots, y_n)$ is observed and $\theta = \theta_1, \ldots, \theta_n$, (the unknown), is to be estimated; two key information needed to achieve this goal (posterior distribution of θ given $\mathbf{X}$) are the prior distribution, $p(\theta)$ from which θ was derived and the likelihood function, $l(\theta \mid \mathbf{X})$, (information from the data which compliments the prior distribution) (Smith and LeSage (2004)). In our case lets assume θ is a vector of continuous variables. With these information on likelihood and prior distributions, the posterior distribution $p(\theta \mid \mathbf{X})$ is said to be proportional to the product of the likelihood and prior probability ased on Baye's theorem (Walsh (2002); Berger (2013)). The constant of proportionality is $\frac{1}{p(\mathbf{X})}$, computed with respect to θ and scales the product to one. The constant $p(\theta)$ can be computed as $p(\mathbf{X}) = \int_\theta \ell(\theta \mid \mathbf{X}).p(\theta)d\theta$ (Berger (2013)).

In our case we wish to estimate the spatial effects which are u_{1i}, u_{2i}, v_{1i} and v_{2i} and variances. The prior $p(\theta = u_{1i}, u_{2i})$ distribution of u_{1i} and u_{2i} are given as Eqution (3.1.21) and Equation (3.1.22) respectively. The values of $d_{1i}, d_{2i}, N(1i)$ and $N(2i)$ are known in our case. The other unknown parameters are v_{1i} and v_{2i} with priors as Equation (3.1.23) and Equation (3.1.24) respectively (Ngesa et al. (2014b)). The unknown variances $\tau_{1i} = \left(k_{1i}, \sigma_{1i}^2\right)$, $\tau_{2i} = \left(k_{2i}, \sigma_{2i}^2\right)$ are given inverse gamma priors (Walsh (2002); Smith and LeSage (2004); Nkurunziza et al. (2010); Berger (2013); Ngesa et al. (2014b)).

As stated in the literature review, estimation is achieved by the MCMC method. In this method, a sample is sequentially sampled from the complete set of conditional

distributions for the parameters. We start by determining the complete set of joint distributions for all unknown parameters. After which the sequential sampling is done or derived. The set of estimates determined from this method converges in the limit to the true (joint) posterior distribution of the unknown parameters(Gelfand and Smith, 1990)).

Now we use the basic Bayesian identity and the priors defined above to compute the conditional posterior distributions.

$$\left.\begin{aligned} p\left(u_{1i}, u_{2i}, v_{1i}v_{2i}, \tau \mid y\right).p\left(y\right) = \quad & p\left(y \mid u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_{1i}, \tau_{2i}\right) \\ & \times p\left(u_{1i}, u_{2i}, v_{1i}v_{2i}, \tau_{1i}, \tau_{2i}\right) \end{aligned}\right\} \tag{3.2.1}$$

where $p\left(.\right)$ is the posterior density involving the y observations.

The posterior joint distribution $p\left(u_{1i}, u_{2i}, v_{1i}v_{2i}, \tau_{1i}, \tau_{2i} \mid y\right)$ is given up to a constant of proportionality by;

$$\left.\begin{aligned} p\left(u_{1i}, u_{2i}, v_{1i}v_{2i}, \pi_i, \tau_{1i}, \tau_{2i} \mid y\right) \propto \quad & p\left(y \mid \lambda_i, \pi_i\right) \times p\left(\lambda_i \mid u_{1i}, v_{1i}, \tau_1\right) \\ & \times p\left(\pi_i \mid u_{2i}, v_{2i}, \tau_2\right) \times p\left(u_{1i} \mid u_{-1i}, k_{1i}\right) \\ & \times p\left(v_{2i} \mid \sigma_{2i}^2\right) p\left(u_{2i} \mid u_{-2i}, k_{2i}\right) \\ & \times p\left(v_{1i} \mid \sigma_{1i}^2\right) \times p\left(k_{1i}\right) \times p\left(k_{2i}\right) \\ & p\left(\sigma_{1i}^2\right) \times p\left(\sigma_{2i}^2\right). \end{aligned}\right\} \tag{3.2.2}$$

With Equation (3.2.1), we establish the posterior marginal distribution of each of the parameters. With this in mind and $\star$ standing for the conditioning arguments $u_{1i}, u_{2i}, v_{1i}v_{2i}, \tau$, we begin with the posterior marginal distribution of u_{1i} as;

$$\left.\begin{aligned} p(u_{1i} \mid \star) &= \frac{p(\lambda_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)}{p(u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)} \\ &\propto p(\lambda_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y) \\ &\propto p(\lambda_i \mid u_{1i}, v_{1i}, \tau_1, \tau_2)\, p(u_{1i}) \\ &\propto exp\left\{-\frac{1}{2\tau_1}\sum_{i=1}^{n}(\lambda_i - \psi_{1i})\right\}.exp\left\{-\frac{1}{2}\sum_{i}\sum_{j\in N(1j)}\left(u_{1i}-u_{1j}\right)^2\right\} \end{aligned}\right\} \tag{3.2.3}$$

where ψ_{1i} represents the covariates, $X\beta$ for region 1.

Therefore the conditional posterior distribution of u_{1i} follows a Normal distribution .

The conditional posterior distribution of u_{2i} can also be calculated as;

$$\left.\begin{aligned} p(u_{2i} \mid \star) &= \frac{p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)}{p(u_{1i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)} \\ &\propto p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y) \\ &\propto p(\pi_i \mid u_{2i}, v_{2i}, \tau_1, \tau_2) \times p(u_{2i}) \\ &\propto exp\left\{-\frac{1}{2}\sum_{i}\sum_{j\in N(j)}\left(u_{2i}-u_{2j}\right)^2\right\}. \end{aligned}\right\} \tag{3.2.4}$$

Therefore the conditional posterior distribution of u_{2i} follows a Normal distribution .

The conditional posterior distribution of v_{1i} can also be computed as;

$$\left.\begin{aligned} p(v_{1i} \mid \star) &= \frac{p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)}{p(u_{1i}, u_{2i}, v_{2i}, \tau_1, \tau_2 \mid y)} \\ &\propto p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y), \\ &\propto p(\lambda_i \mid u_{1i}, v_{1i}, \tau_1, \tau_2)\, p(v_{1i}), \\ &\propto exp\left\{-\frac{1}{2\tau}\sum_{i=1}^{n}(\lambda_i - \psi_{1i})^2\right\}.exp\left\{-\frac{1}{2\sigma_{1i}^2}\sum_{i=1}^{n}v_{1i}^2\right\}. \end{aligned}\right\} \tag{3.2.5}$$

where ψ_{2i} represents the covariates producing π_i.

The conditional posterior distribution of v_{2i} can be computed as;

$$
\left.
\begin{aligned}
p(v_{2i} \mid \star) &= \frac{p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y)}{p(u_{1i}, u_{2i}, v_{1i}, \tau_1, \tau_2 \mid y)} \\
&\propto p(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y), \\
&\propto p(\lambda_i \mid u_{1i}, v_{2i}, \tau_1, \tau_2)\, p(v_{2i}), \\
&\propto exp\left\{-\frac{1}{2\tau}\sum_{i=1}^{n}(\lambda_i - \psi_{2i})^2\right\}.exp\left\{-\frac{1}{2\sigma_{1i}^2}\sum_{i=1}^{n} v_{2i}^2\right\}.
\end{aligned}
\right\} \tag{3.2.6}
$$

The conditional posterior distribution of τ can be computed as;

$$
\left.
\begin{aligned}
p(\tau \mid \star) = \quad & p(\tau \mid \alpha, \delta)\, p(\tau) \\
& \propto (\tau)^{-(\alpha+1)}\, exp\left(-\frac{\delta}{\tau}\right),
\end{aligned}
\right\} \tag{3.2.7}
$$

where α and δ are hyper parameters of the gamma distribution.

The advantage of Bayesian over classical statistics (MLE) is, confounding effects of latent parameters are eliminated leading to better estimate of the posterior distribution of some unknown parameter from a joint marginal posterior distribution. Another, more practical approach is the addition of intangible information which will otherwise be hidden in data. Not all characteristics can be measured numerically especially with non-infectious disease modeling. Aside the nature of the population at risk, one core characteristic that introduces spatial effect is the geographical locality of the population. Enumerating such an influential indicator will be difficult to achieve. However, the addition of collected data will cater for this defect taking into consideration the fact that the population under study is partitioned into one hundred and thirty-eight units. The MCMC approach can be extended to a multivariate distribution where a sequence of observations can be generated by a sampling method that employs MCMC algorithm, especially when direct sampling fails. Two of such methods are Gibbs sampling and Metropolis–Hastings algorithm which was used in the analysis of Models (3.1.28,3.1.29,3.1.30 and 3.1.31).

Chapter 4

Results and Discussion

Annual diabetes cases in Ghana falls within the brackets of one (1) to 29,474 with the former being one of the newly created districts; North Tongu in the Volta Region and the latter being Tema-Kpone-Akatamanso district in the Greater Accra region. Diabetes cases are presumably low in the northern part of the country with the opposite occurring as one travels down the South. A possible explanation could be the vast variation in population. This pattern or occurrence is not different when one plots the disease map for the rate of disease occurrence.

Model estimation was carried out using a Bayesian approach with every parameter being assigned prior distributions. To be precise, a non informative normal prior was assigned to the offset parameter, α_0 whiles the variance parameters are assigned inverse gamma distributions. The thesis was carried out under the assumption that covariates are not available. WinBugs version 1.4 was used in the implementation (Spiegelhalter et al. (2003)) phase. A double chain of MCMC iterations of 120,000 were ran with initial of 10,000 left out as the burn-in period and then every tenth sample value considered for arriving at the convergence of the estimates the remaining 11,000 samples. The decision on convergence was arrived at based on the behavior of our trace plots and auto-correlation plots of the MCMC output. The models were compared using

the Deviance Information Criterion (DIC) as proposed by (Spiegelhalter et al. (2003)). The best fitting model,i.e, Model (3.1.31) was usd in the analysis.

The map helps to throw more light on the state of disease in the area under discussion. With reference to this map, policy makers and stakeholders can formulate pragmatic policies to tackle diseases in each district of Ghana thereby resulting in proper allocation of resources in countering these diseases.

In this work, no explanatory variables were considered except for the fact that the relative risk was made to depend on only spatial properties, i.e. correlated (u_{1i}) and uncorrelated (v_{1i}) spatial properties. Although the mean values of u_1 falls between $(-9.072, 9.306)$ but the 95 Credible Interval (CI) values are all in the positive range which depicts a positive relationship with the relative risk. The same case follows for v_1, where the mean values falls in the range $(-1.761, 2.986)$, the 95% CI are positive values.

The introduction of the spatially varying probability of under-reporting for each district reduced the DIC of Model 3.1.28 (Model without under-reporting probability) by a wide margin of more than 1000 (Table 4.1). This makes the Poisson model with the under-reporting a better alternative to the Poisson model without the under-reporting parameters.

In Figure (4.4), which is the disease map for the probability of under-reporting, it can be seen from the legends that, most of the values fall between $(0.1 - 0.9)$ with few falling below and above 0.1 and 0.9 respecteively

The mean values of the correlated part, u_2, of the spatial probability spans from -107.9 to $52, 25$ although it has more positive values as its Credible Intervals. A careful analysis of Table (4.1) shows that the magnitude of the node is mostly influenced by the the uncorrelated part v_2 of the spatial variation:- as the spatial probability node increases with increase in (v_2). The 95% credible interval of v_2 ranges from $(-13.46, 630.5)$ with most of the figures been positive, signifying positive relationship between the rel-

ative risk of diabetes and the unstructured spatial effect for under-reported cases, v_2. Although the 95% credible interval which ranges from $(-9.809, 16.87)$, most of the range was in the positive range signifying positive relationship between the structured spatial effect for under-reported cases, u_2 and the relative risk. From Appendix B, it can be seen that the Monte Carlo (MC) error values were all less that 5% for each of the nodes in all units signifying a clear case of convergence.

In Figure (4.1), it is seen from the legends that, majority of the districts are in the lower risk areas with the very few at higher risk. The lowest risks-prone areas are predominantly found in some of the three northern regions. Some of which include Wa West, Yendi and Bongo. This is clearly buttressed by the plot of the lower credible interval disease map, Figure (4.2) where 137 out of 138 districts are below the 10 mark and are all in the northern part of the country. On the other hand, the upper credible interval map from Figure (4.3) also has most of the districts above the 10 mark looking at the legend, with only 2 of them falling below the 0.1 mark. All of the districts falling in this domain are all in the southern part of the country.

The districts with the highest risk are found in the southern part of the country with examples as: Komenda-Edina-Eguafo-Ebirem, Obuasi, Akwapim North and Amansi West hitting beyond the 20 mark.

Table 4.1: Comparison of Count Models in Ghana

	Model			
	1	2	3	4
α_0	$-1.537\,(-1.8,-1.28)$	$-0.6671\,(-0.8246,-0.6866)$	$-0.2721\,(-0.5113,-0.05194)$	$-0.346\,(-0.6345,-0.03041)$
$\sigma^2_{u_1}$	$6.4\,(4.243,2.41)$	$7.853\,(3.75,12.72)$	$6.41\,(4.376,8.22)$	$6.595\,(3.729,10.476)$
$\sigma^2_{v_1}$	$2.026\,(1.642,2.41)$	$1.859\,(1.357,2.296)$	$1.275\,(0.2702,1.743)$	$6.595\,(3.729,10.476)$
$\sigma^2_{u_2}$	*NA*	$146.7\,(0.1498,393.6)$	*NA*	$147.2\,(0.1518,413.7)$
$\sigma^2_{v_2}$	*NA*	*NA*	$16.45\,(11.9,21.03)$	$17.06\,(9.326,25.306)$
pD	153.187	-970.3210	-10317.60	-19856.100
DIC	1383.370	258.993	-9093.67	-10631.400

Figure 4.1: Diabetes relative map of Ghana.

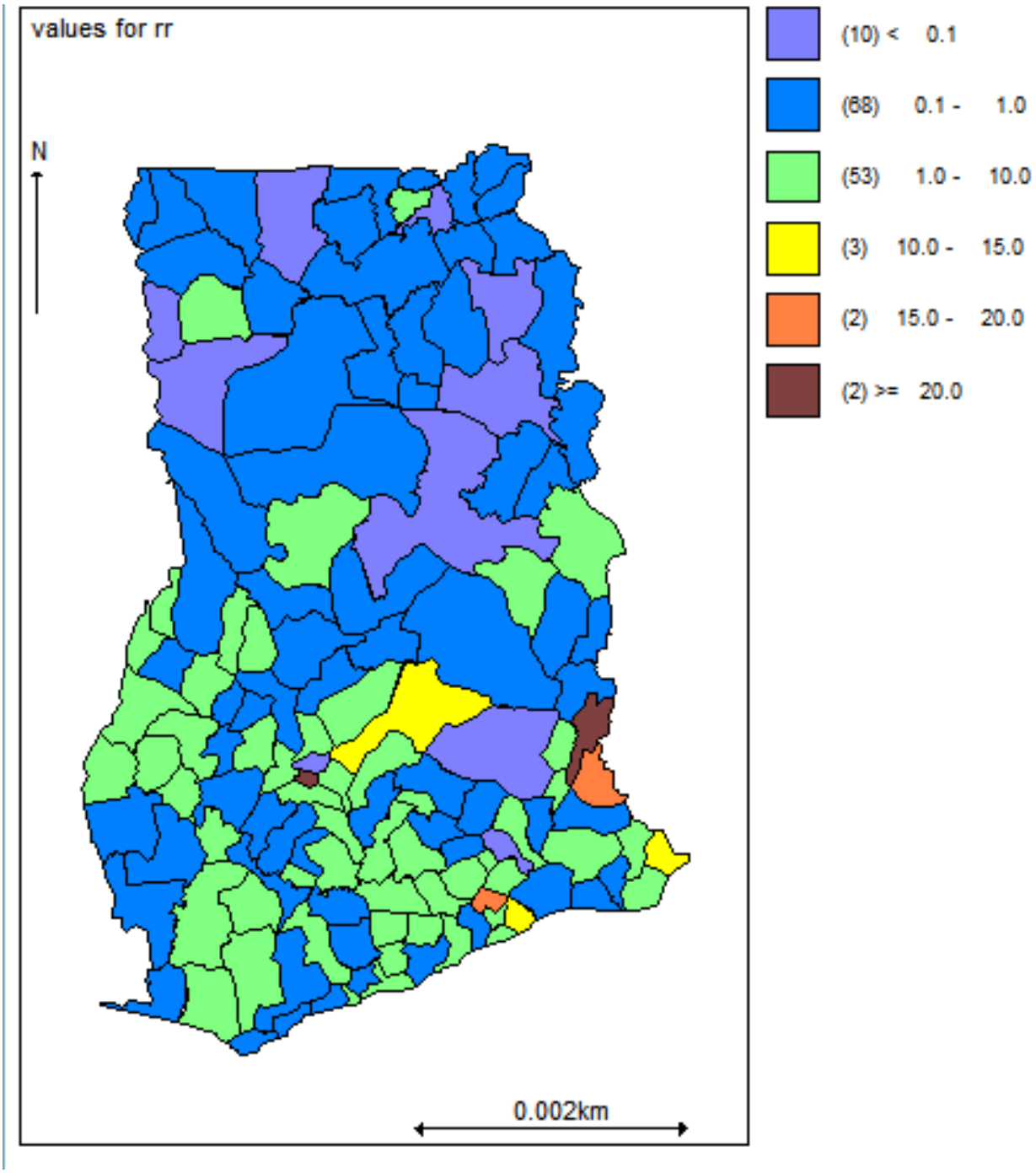


Figure 4.2: The map of 2.5% credible Interval.

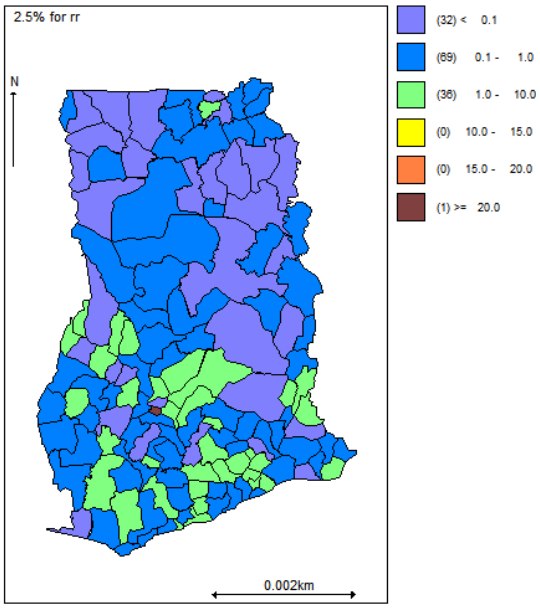


Figure 4.3: The map of 97.5% credible Interval

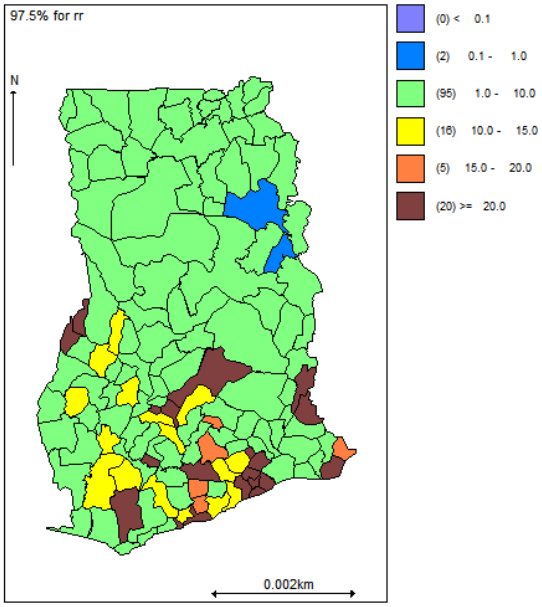
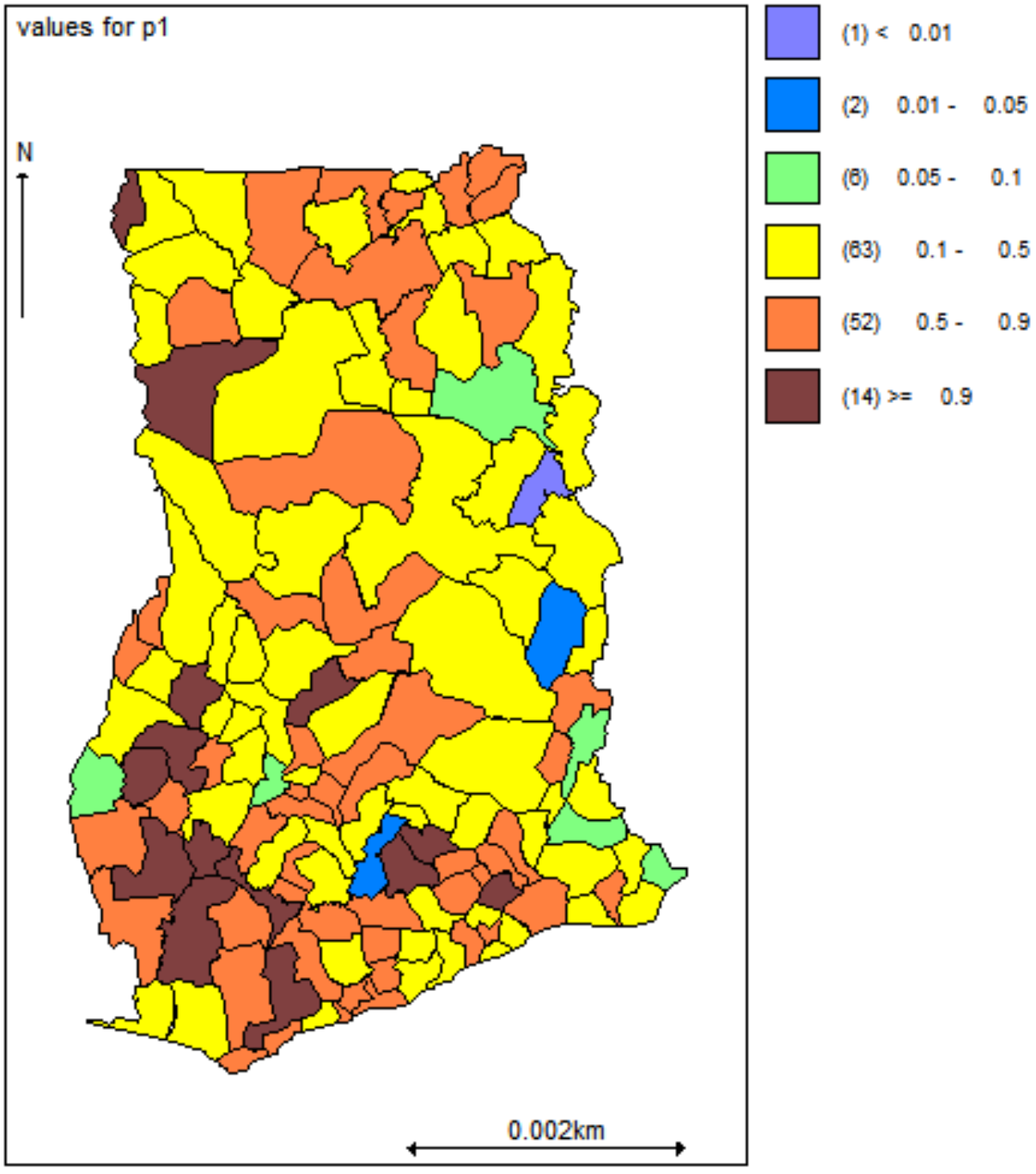


Figure 4.4: Varying probability of under-reporting in Ghana.



Chapter 5

Conclusions and Recommendations

5.1 Introduction

Disease mapping is gradually gaining grounds in the field of epidemiology for the graphical display of the relative risk of some disease in a given geographical area. However, incorporating under-reporting in econometric models in order to reduce bias when estimating such models is a dominant problem. Under-reporting is mostly present because, given a study population, there is always the chance that some will not report to the recording authorities. Under-reporting also come about when one is subjected to tedious and complex recording mechanisms. Occurrence of such cases of under-reporting can create bias in estimation and also blurr the true state of some disease cases in a given population.

In this paper, a model is devloped for count data taking into consideration that it is lagged with under-reporting and a Bayesian method is applied in estimating the identified parameters. A Poisson regression model for the count data is proposed and with the assumption of independence relaxed, the relative risk is made to depend on some spatial components, proposed by (Besag and Newell, 1991b). Under-reporting on the other hand is captured by a logit probability and made to vary spatially from one district

to the other. This addition is an important contribution in the study area. We further compute and estimate the spatial effects of the relative risk and the under-reporting probability. Unobserved variables which contributes immensely to the relative risk are treated as latent variables and employing methods in Gibbs sampling, the non-reported variables or incidences are added, arriving at inferences on the model parameters. In this paper, covariates are not included thereby flushing out the case of estimating state-depedent variables. However, Bayesian estimation approach is used adding additional value to the spatial parameters been estimated.

The proposed model was then applied to data collected over a period of one year on monthly basis. Analysis of the model in relation to data clearly shows that, in the presence of spatial auto-correlation and particularly under-reporting, the proposed model produces a very good fit to the diabetes data. In this work, we model spatially varying count data with cases of under-reporting. This we do by first of all, using a Poisson distribution to model the count data and the under-reporting represented by a binomial probability with a logit as a link function. The average parameter of the distribution is a product of the relative risk, λ and the expected count, E. With the assumption of independence relaxed, the relative risk, λ can be made to depend on contribtuing covariates. In this thesis, the contributing factors are solely correlated spatial effect, u_i and uncorrelated spatial effect, v_i.

The third specific objective which is developing a disease map from the estimated relative risk is achieved by a method likend to model validation. Using WinBugs 1.4, the achieved model in specific objective two (2) is sitmulated using real data and the map generated after convergence was achieved. In achieving required results, the spatially structured random effects were captured by the usual Conditional Auto-regressive (CAR) model proposed by (Besag and Newell, 1991a) with parameters and hyper-parameters assigned non-informative priors. The under-reporting parameter is a probability which is allowed to vary spatially over the districts. This model is called

spatial-pi-model. The model is a suitable alternative to most convolution models as there is an advantage of capturing under-reported cases.

5.2 Conclusions

Quite a number of models were compared to the proposed model; some of which include, the logistic and some convolution models. The spatial-pi-model (new model) has a lesser DIC than the rest of the models making it a better choice. The spatial-pi-model was used in producing diabetes prevalence smoothed maps for districts of Ghana which can be used to advice policy makers on the dispensing of resources to fight diabetes diseases. The spatially varying probability of under-reporting corrects the lapses created by complex and tedious data collection mechanisms. This position is confirmed by earlier works done by (Winkelmann and Zimmermann, 1993; Yang et al., 2010).

Although Tema-Kpone-Akatanmanso recorded the highest number of cases in Ghana, those in Techiman Municipal, Fanteakwa, Agona East, Ejura-Sekyedumase and Bosomtwi were found to be at the highest risk of contracting diabetes. These results are in the same direction with the findings of (Darkwa, 2011) who identifies the Central, Brong Ahafo and the Ashanti Regions of Ghana as some of the regions with diabetes cases which are higher than the world average; at least one of the districts named above can be found in at least one of these three regions. Quaye et al., 2015; Darkwa, 2011 in their works identified the continual growth of diabetes cases in Ghana with Danquah et al., 2012; Darkwa, 2011 identifying some of the factors associated with the disease as the consumption of carbohydrates and fatty foods, urbanization and less exercising among the population. These findings make a lot of sense when subjected to the five populations identified. This is because the main or traditional foods of the inhabitants are basically made from cassava and maize even though beans, yam and maize are eaten across Ghana (Frank et al., 2014). Also, one of the reasons for the high risk

could be as a result of urbanization which is in line with the findings of (Al-Lawati and Jousilahti, 2004), where urbanization was accompanied with increase in income and subsequently increase in car ownership and consumption of fatty foods. There is a visible absence of recreational and aerobic activities in the identified populations (Darkwa, 2011).

On the other side of the coin are Upper Manya Krobo, Ga East and La-Nkwanta, Savelugu-Nanton, Kpandai and Nanumba North and South districts, which recorded the lowest risk in ascending order. In the case of the first two, this could be associated to the availability of health resources in the battling of the menace as fairly well equipped health facilities are available (Darkwa, 2011). The other districts however are found in the northern part of Ghana. In this part of the country, their main staple food is made from millet, sorghum, guinea corn and groundnut with visibly small or no protein contents. This extends to the fact that, carbohydrates and fatty foods are kept at the barest minimum Frank et al. (2014). Also, inhabitants of these region trek miles from their homes to farms and there is virtually no means of transportation. In the cases where transportation is available, it comes at a cost. This is in line with many research works where exercising has been identified to be one of the factors that retrogresses diabetes (Al-Lawati and Jousilahti, 2004; Darkwa, 2011; Frank et al., 2014.)

5.3 Recommendations

In this work, validation was not based on covariates as the main idea was to identify and correct spatial effects in under-reporting cases in each district. Disease maps are very helpful in the area disease combat. Policy makers make reference to these (disease maps) when allocating our scarced resources in controlling diabetes in Ghana.

For the sake of future works and recommendation, we propose an extension in the effect of looking at multivariate domain where multiple diseases are known to ex-

ist in each geographical setting. Also, some assumptions like the exclusion of over-dispersion should be relaxed and incorported in future research works by using negative binomial distribution instead of Poisson distribution.

Bibliography

Aitkin, M. (1991). Posterior bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 111–142.

Al-Lawati, J. A. and Jousilahti, P. J. (2004). Prevalence and 10-year secular trend of obesity in oman. *Saudi medical journal*, 25(3):346–351.

Almani, S. A., Memon, A. S., Memon, A. I., Shah, I., Rahpoto, Q., and Solangi, R. (2008). Cirrhosis of liver: Etiological factors, complications and prognosis. *J Liaquat Uni Med Health Sci*, 7(2):61–6.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Besag, J. and Newell, J. (1991a). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 143–155.

Besag, J. and Newell, J. (1991b). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 143–155.

Cressie, N. (1993). Statistics for spatial data: Wiley series in probability and statistics. *Wiley-Interscience, New York*, 15:105–209.

Danquah, I., Bedu-Addo, G., Terpe, K.-J., Micah, F., Amoako, Y. A., Awuku, Y. A., Dietz, E., van der Giet, M., Spranger, J., and Mockenhaupt, F. P. (2012). Diabetes

mellitus type 2 in urban ghana: characteristics and associated factors. *BioMed Central*, 210(3).

Darkwa, S. (2011). Prevalence of diabetes mellitus and resources available for its management in the cape coast metropolis. *ISABB Journal of Health and Environmental Sciences*, 1(1):1–7.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, 7(4):247–252.

Draper, N. R. and Smith, H. (1998). Applied regression analysis . new york: John willey & sons.

Feller, W. (1968). *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons London-New York-Sydney-Toronto.

Frank, L. K., Kröger, J., Schulze, M. B., Bedu-Addo, G., Mockenhaupt, F. P., and Danquah, I. (2014). Dietary patterns in urban ghana and risk of type 2 diabetes. *British Journal of Nutrition*, 112(01):89–98.

Gamado, K. M., Streftaris, G., and Zachary, S. (2014). Modelling under-reporting in epidemics. *Journal of mathematical biology*, 69(3):737–765.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.

Gibbons, C. L., Mangen, M.-J. J., Plass, D., Havelaar, A. H., Brooke, R. J., Kramarz, P., Peterson, K. L., Stuurman, A. L., Cassini, A., Fèvre, E. M., et al. (2014). Measuring underreporting and under-ascertainment in infectious disease datasets: a comparison of methods. *BMC Public Health*, 14(1):1.

Gilks, W. R. (2005). *Markov chain monte carlo*. Wiley Online Library.

IDF (2014). Diabetes scorecard in Ghana;http://www.idf.org/regions/africa. *IDF Press*.

Koch, T. (2005). *Cartographies of disease: maps, mapping, and medicine*. Esri Press Redlands, CA.

Moore, D. A., Carpenter, T. E., et al. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic reviews*, 21(2):143–161.

Moraga, P. and Lawson, A. B. (2012). Gaussian component mixtures and car models in bayesian disease mapping. *Computational Statistics & Data Analysis*, 56(6):1417–1433.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Neubauer, G., Djuraš, G., and Friedl, H. (2016). Models for underreporting: A bernoulli sampling approach for reported counts. *Austrian Journal of Statistics*, 40(1&2):85–92.

Ngesa, O., Achia, T., and Mwambi, H. (2014a). A flexible random effects distribution in disease mapping models. *South African Statistical Journal*, 48(1):83–93.

Ngesa, O., Mwambi, H., and Achia, T. (2014b). Bayesian spatial semi-parametric modeling of hiv variation in kenya. *PloS one*, 9(7):e103299.

Nkurunziza, H., Gebhardt, A., and Pilz, J. (2010). Bayesian modelling of the effect of climate on malaria in burundi. *Malaria journal*, 9(1):1.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, volume 698. John Wiley & Sons.

Quaye, E. A., Amporful, E. O., Akweongo, P., and Aikins, M. K. (2015). Analysis of the financial cost of diabetes mellitus in four cocoa clinics of ghana. *Value in Health Regional Issues*, 7:49–53.

Ripley, B. D. (1977). Modelling spatial patterns. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 172–212.

Rodrıguez, G. (2007). Poisson models for count data. *Lecture Notes on Generalized Linear Models, http://data. princeton. edu/wws509/notes*.

Smith, T. E. and LeSage, J. P. (2004). A bayesian probit model with spatial dependencies. *Advances in econometrics*, 18:127–160.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). Winbugs user manual.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Research Report, 98-009.

Starkweather, J. (2011). Sharpening occam's razor: Using bayesian model averaging in r to separate the wheat from the chaff. *Benchmarks RSS Matters*.

Tango, T. (2010). *Statistical methods for disease clustering*. Springer Science & Business Media.

Walsh, B. (2002). Introduction to bayesian analysis. *Lecture notes for EEB 596z*.

Wartenberg, D. (1999). Using disease-cluster and small-area analyses to study environmental justice.

WHO (2014). Risk predictive modelling for diabetes and cardiovascular diseas. *Bulletin of the World Health Organization*, 51(1).

Winkelmann, R. (1996). Markov chain monte carlo analysis of underreported count data with an application to worker absenteeism. *Empirical Economics*, 21(4):575–587.

Winkelmann, R. and Zimmermann, K. F. (1993). *Poisson-logistic regression*. Volkswirtschaftl. Fakultät d. Ludwig-Maximilians-Univ. München.

Yang, S., Zhao, Y., and Dhar, R. (2010). Modeling the underreporting bias in panel survey data. *Marketing Science*, 29(3):525–539.

Ye, F. and Lord, D. (2011). Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. *Transportation Research Record: Journal of the Transportation Research Board*, (2241):51–58.

Zayeri, F., Salehi, M., and Pirhosseini, H. (2011). Geographical mapping and bayesian spatial modeling of malaria incidence in sistan and baluchistan province, iran. *Asian Pacific journal of tropical medicine*, 4(12):985–992.

Appendix A

WinBugs codes

```
model{
for (i in 1:N)
#N=138
{
y[i]~dpois(mu1[i])

mu1[i]<-mu[i]*p1[i]
p1[i]<-min(1,max(0.001,p[i]))

log(mu[i])<-log(e[i]) + log(rr[i])

log(rr[i])<- aph0 +v1[i] +u1[i]
v1[i]~dnorm(0,tau.v1)

logit(p[i])<- u2[i]+ v2[i]
v2[i]~dnorm(0,tau.v2)
}
#Prior
```

```
for(k in 1:sumNumNeigh)
{ weights[k]<-1

}
u1[1:N]~car.normal(adj[],weights[], num[],tau.u1)
u2[1:N]~car.normal(adj[],weights[], num[],tau.u2)

#PRIOR DISTRIBUTIONS (our model has no fixed effects
#components, that is the reason why we don't have betas)
tau.v1~dgamma(0.001,0.001)
tau.u1~dgamma(0.001,0.001)
tau.v2~dgamma(0.001,0.001)
tau.u2~dgamma(0.001,0.001)
aph0~dflat()
sigma.u1 <- sqrt(1 /tau.u1) # standard deviation of u1
sigma.v1<- sqrt(1 / tau.v1) # standard deviation of v1
sigma.u2<- sqrt(1 / tau.u2) # standard deviation of u2
sigma.v2<- sqrt(1 / tau.v2) # standard deviation of v2

}

#Data


#Initialization
```

Spatial Variation of Diabetes Cases in some Districts in Ghana

ID	District Name	Mean (Credible Intervals)	MC Error
1	Adansi North	0.1137 (0.0936, 0.1359)	0.0151
2	Adansi South	0.7702 (0.7155, 0.8269)	0.0001
3	Afigya Kwabre	1.033 (0.9737, 1.0950)	0.0004
4	Ahafo-Ano North	0.3203 (0.2827, 0.3609)	0.0004
5	Ahafo-Ano South	1.8160 (1.7170, 1.9120)	0.0003
6	Amansie Central	0.0669 (0.0528, 0.0832)	0.0006
7	Amansie West	0.7514 (0.7077, 0.7976)	0.0001
8	Asante-Akim Central	0.2223 (0.1906, 0.2558)	0.0003
9	Asante-Akim South	1.1680 (1.1050, 1.2350)	0.0002
10	Asante Mampong	0.1070 (0.0875, 0.1278)	0.0004
11	Atwima Kwanoma	0.7634 (0.7250, 0.8042)	0.0001
12	Atwima Mponua	0.0352 (0.0240, 0.0482)	0.0003
13	Atwima-Nwabiagya	0.601 (0.5495, 0.6568)	0.0009
14	Bekwai	5.5110 (5.2960, 5.7350)	0.0004
15	Bosome-Freho	0.2010 (0.1769, 0.2258)	0.00153
16	Bosomtwi	25.2500 (24.6900, 25.8300)	0.0002
17	Ejisu-Juaben	0.0269 (0.0245, 0.0294)	0.0002
18	Ejura-Sekyedumase	6.3670 (6.1740, 6.5670)	0.0016
19	Kumasi	0.0985 (0.0833, 0.1145)	0.0002
20	Kwabre East	5.0950 (4.8780, 5.3180)	0.0016
21	Obuasi	1.1970 (1.1370, 1.2570)	0.0005
22	Offinso	2.4940 (2.3940, 2.5970)	0.0007
23	Offinso North	0.8203 (0.7574, 0.8851)	0.0005
24	Sekyere Central	0.9520 (0.8864, 1.0190)	0.0005
25	Sekyere South	0.4438 (0.4015, 0.4872)	0.0003

ID	District Name	Mean (credible Intervals)	MC Error
26	Asutifi	0.0616 (0.0473, 0.0778)	0.0001
27	Atebubu Amanten	0.6130 (0.5752, 0.6501)	0.0003
28	Berekum	3.86 (3.7050, 4.0210)	0.0012
29	Dormaa East	1.6520 (1.5620, 1.7480)	0.0007
30	Dormaa Municipal	1.0050 (0.9364, 1.0750)	0.0005
31	Jaman North	0.5210 (0.4677, 0.5762)	0.0004
32	Jaman South	0.1743 (0.1537, 0.1967)	0.0002
33	Kintampo North	0.4676 (0.4284, 0.5079)	0.0003
34	Kintampo South	0.0512 (0.0383, 0.0662)	0.0001
35	Nkoranza North	1.1820 (1.1310, 1.2330)	0.0004
36	Pru	0.0359 (0.0249, 0.0488)	0.0009
37	Sene	0.08592 (0.06535, 0.1095)	0.0001
38	Sunyani Municipal	0.0207 (0.0087, 0.0380)	0.0001
39	Sunyani West	1.2920 (1.2350, 1.3490)	0.0004
40	Tain	1.8750 (1.7730, 1.9830)	0.0008
41	Tano North	0.1798 (0.1545, 0.2067)	0.0002
42	Tano South	0.1866 (0.1666, 0.2062)	0.0001
43	Techiman Municipal	1.1000 (1.0400, 1.1620)	0.0005
44	Wenchi	2.2360 (2.1350, 2.3410)	0.0008
45	Abura-Asebu	0.3784 (0.3473, 0.4116)	0.0002
46	Agona East	0.0983 (0.0791, 0.1198)	0.0001
47	Agona West	0.8027 (0.758, 0.8476)	0.0003
48	Ejumako-Enyam	5.3780 (5.2310, 5.5290)	0.0012
49	Asikuma-Odoben	0.4243 (0.3946, 0.4550)	0.0001
50	Assin North	0.1651 (0.1436, 0.1888)	0.0002

ID	District Name	Mean (credible Intervals)	MC Error
51	Assin South	6.8140 (6.6410, 6.9880)	0.001108
52	Awutu senya	1.1780 (1.1100, 1.2470)	0.0005
53	Cape Coast	0.8339 (0.7814, 0.8894)	0.0004
54	Ekumfi	0.0151 (0.0101, 0.0210)	0.0004
55	Gomoa East	4.7010 (4.5490, 4.8550)	0.0011
56	Gomoa West	2.4690 (2.3660, 2.5730)	0.0008
57	Komenda Edna	0.0940 (0.07499, 0.1005)	0.0002
58	Mfantsipim	0.1001 (0.0809, 0.1207)	0.0002
59	Twifo Aheman	0.0399 (0.0359, 0.0441)	0.0003
60	Upper Denkyira E.	3.309 (3.2140, 3.4050)	0.0005
61	Upper Denkyira W.	0.6113 (0.5701, 0.6534)	0.0003
62	Akwapim N.	2.901 (2.8050, 3.0010)	0.0002
63	Akwapim South	1.2540 (1.1920, 1.3180)	0.0005
64	Akyemansa	1.2370 (1.1800, 1.2930)	0.0001
65	Asuogyamang	0.0489 (0.0344, 0.0662)	0.0009
67	Atiwa	0.7965 (0.7770, 0.8164)	0.0003
68	Birim Central	0.0722 (0.0601, 0.0848)	0.0007
69	Birim South	0.2747 (0.2462, 0.3051)	0.0005
70	East Akim	1.0820 (1.0380, 1.127)	0.0003
71	Fanteakwa	0.3308 (0.2833, 0.3825)	0.0001
72	Kwabibirem	0.1451 (0.1236, 0.1684)	0.0001
73	Kwahu East	0.4894 (0.4417, 0.5385)	0.0009
74	Kwahu North	0.4894 (0.4417, 0.5385)	0.0002
75	Kwahu South	0.0338 (0.0265, 0.0419)	0.0003

ID	District Name	Mean (credible Intervals)	MC Error
76	Kwahu West	0.2581 (0.2281, 0.2896)	0.0002
77	Lower Manya Krobo	0.0487 (0.0400, 0.0633)	0.0008
78	New Juabeng	0.0537 (0.0378, 0.0718)	0.0003
79	Suhum Kraboa-Coaltar	0.1053 (0.0880, 0.1236)	0.0002
80	Upper Manya Krobo	0.0029 (0.0006, 0.0074)	0.0004
81	West Akim	0.0812 (0.0650, 0.0991)	0.0006
82	Yilo Krobo	0.0712 (0.0571, 0.0875)	0.0002
83	Accra Metro	0.0634 (0.0479, 0.0817)	0.0010
84	Adentan	0.2746 (0.2567, 0.2929)	0.0001
85	Ashaiman	0.0475 (0.0355, 0.0621)	0.0001
86	Damgme (Ada east)	0.1746 (0.1458, 0.2057)	0.0003
87	Damgme(Shai Osudoku)	0.1990 (0.1765, 0.2214)	0.0001
88	Ga East	0.0057 (0.0028, 0.0096)	0.0001
89	Ga South	0.1185 (0.0988, 0.1399)	0.0001
90	Ga west	0.6167 (0.5789, 0.6551)	0.0001
91	Ledzokuku	0.2006 (0.1701, 0.2329)	0.0001
92	Tema Kpone	0.2006 (0.1701, 0.2329)	0.0002
93	Bole	1.025 (0.9622, 1.089)	0.0002
94	Bunkpurugu	0.0325 (0.0204, 0.0468)	0.0002
95	Central Gonja	0.1677 (0.1399, 0.1978)	0.0002
96	Chereponi	0.4652 (0.4248, 0.5081)	0.0003
97	East Gonja	0.4328 (0.3921, 0.4770)	0.0002
98	East Mamprusi	0.0258 (0.0166, 0.0367)	0.0005
99	Gushiegu	0.0561 (0.0438, 0.0707)	0.0009
100	Karaga	0.1676 (0.1405, 0.1955)	0.0002

ID	District Name	Mean (credible Intervals)	MC Error
101	Kpandai	0.0877 (0.0685, 0.1089)	0.0003
102	Nanumba N.	0.0162 (0.0071, 0.0294)	0.0003
103	Nanumba S.	0.0477 (0.0293, 0.0712)	0.0008
104	Saboba	0.0505 (0.0370, 0.0661)	0.0009
105	Savelugu-Nanton	0.6500 (0.5876, 0.7152)	0.0002
106	Sawla-Tuna	0.0057 (0.0017, 0.0122)	0.0002
107	Tamale-Savelugu	0.0167 (0.0079, 0.0284)	0.0009
108	Tolon Kunbugu	0.6261 (0.5795, 0.6749)	0.0002
109	West Gonja	1.8650 (1.7990, 1.9350)	0.0010
110	West Mamprusi	2.2590 (2.1250, 2.3950)	0.0005
111	Yendi+Miom	0.4171 (0.3788, 0.4576)	0.0005
112	Zabzugu+Tatale	0.3679 (0.3347, 0.4027)	0.0009
113	Bawku+Binduri	2.4110 (2.3010, 2.5220)	0.0003
114	Bawku West	0.7513 (0.7167, 0.7872)	0.0005
115	Bolgatanga	1.0320 (0.9690, 1.0980)	0.0009
116	Bongo	0.7243 (0.6545, 0.7970)	0.0008
117	Builsa	0.0078 (0.0034, 0.0137)	0.0002
118	Garu	0.6401 (0.5918, 0.6904)	0.0003
119	Kasena-Nankana	0.5065 (0.4661, 0.5490)	0.0005
120	Kasena-Nankana W.	0.7242 (0.6911, 0.7585)	0.0006
121	Talensi+Nabdam	0.2239 (0.1921, 0.2579)	0.0004
122	Jirapa	0.3890 (0.3497, 0.4307)	0.0003
123	Lambussi-Kami	0.2280 (0.2009, 0.2557)	0.0003
124	Lawra+Nadowli	0.1488 (0.1255, 0.1740)	0.0002
125	Nadowli Kaleo	2.2150 (2.1200, 2.315.0)	0.0002

ID	District Name	Mean (credible Intervals)	MC Error
126	Sissala East	$0.0679\,(0.0548, 0.0827)$	0.0003
127	Sissala West	$0.4128\,(0.3732, 0.4542)$	0.0001
128	Wa	$0.8293\,(0.7748, 0.8845)$	0.0002
129	Wa East	$0.4949\,(0.4556, 0.5350)$	0.0007
130	Wa West	$0.4061\,(0.3774, 0.4354)$	0.0001
131	Adaklu+Agortime	$0.5365\,(0.4820, 0.5916)$	0.0003
132	Akatsi	$0.8333\,(0.7819, 0.8841)$	0.0004
133	Biakoye	$2.8180\,(2.7320, 2.9060)$	0.0003