International Journal of Statistics and Probability; Vol. 6, No. 3; May 2017
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Estimating Disease Risk of Diabetes Cases in the Presence of Underreporting

Oti-Boateng Emmanuel[1]

[1] Pan African University Institute of Basic Sciences, Technology and Innovation, Kenya

Correspondence: Oti-Boateng Emmanuel, Pan African University Institute of Basic Sciences, Technology and Innovation, Kenya. Tel: 233-264-885-117. E-mail: oti-boateng@aims-cameroon.org

Received: January 5, 2017 Accepted: March 28, 2017 Online Published: April 23, 2017

doi:10.5539/ijsp.v6n3p188 URL: https://doi.org/10.5539/ijsp.v6n3p188

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 the case of count data. In particular, we allow under-reporting to vary spatially by regions through a probability captured by a binomial distribution. Count data mostly comes with a common property, which is the variance is greater than mean. When this happens, the recommended distribution is Negative Binomial (NB) instead of the usual Poisson distribution. The spatial variations of the disease were divided into correlated and uncorrelated parts. When a Negative Binomial 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 of contribution coming from the uncorrelated part. The model obtained was applied to diabetes data in Ghana. Disease maps for the diseases were also developed for Ghana. These maps are critical and informative to policy makers when coming up with preventive mechanisms in the face of scarce resources.

Keywords: spatial analysis, disease mapping, Markov chain Monte Carlo, Bayesian statistics

1. Introduction

One of the greatest challenges hindering the progress of Africa is non infectious diseases; which put strain on the tax payer and also depletes our human resources at an alarming rate. A good example of such diseases is diabetes. A fact sheet published in November 2016 by the World Health Organization (WHO), has it that, the number of people with diabetes has increased from 108 million in 1980 to 422 million in 2014. It is also estimated that, diabetes prevalence in adults has risen from 4.7 percent in 1980 to 8.5 percent. Diabetes is a chronic disease that comes about as a result of the inability for the pancreas to produce enough insulin or when the body is unable to effectively utilise the insulin produced by the pancreas. It is a leading cause of death in developing countries of which Ghana is included. In 2014, the International Diabetes Federation (IDF) estimated the number of people living with the disease to be 450,000 raising huge concerns among tax payer as to the remedies being put in place to check the menace.

The estimations above have necessitated the need to apply new and modern methods in solving this menace. Statisticians and mathematicians have responded to this by employing regression models in connecting count data to some variables proven to have a significant effect on the disease. As humans operate in space, it is impossible to separate environmental factors to these diseases. This was confirmed by John Snow, when he connected a certain borehole to cholera deaths in 1854 in London. This pioneered the use of geographically-coded data in modelling diseases employing a method referred to as spatial modelling and it has vast applications in the estimating the relative risk of small areas of a geographical location. With this information at hand, one is able to determine the relative risk of exposure thereby serving as a direction to policy makers.

Many papers have been published in this field. Some of which are (Ugarte, Ibez, & Militino, 2006). They touched on the different techniques one can use when faced with modelling risk in mortality data. They also presented a list of smoothing methods based on Poisson inference that estimate mortality rates and ratios better. In their work, the over-dispersion (which comes about as a result of spatial autocorrelation, unstructured heterogeneity or a combination of the two) was identified and accounted for by incorporating random effects into the models. Also, (Waller and Carlin and Xia & Gelfand, 1997) did a remarkable work on disease mapping where some regional mortality and morbidity cases were mapped. In their work, they were able to identify the fact that, Bayes and empirical Bayes methods help to reduce or eliminate the instability of estimates in low-population areas while maintaining their geographical resolution. Based on this knowledge, they extended their work by incorporating temporal effects and spatio-temporal interactions into their model and fitting their data using Markov chain Monte Carlo (McMC).

(Gamado, Streftaris, & Zachary, 2014) also worked on modelling under-reporting in epidemics by considering the s-

tochastic Markov SIR epidemic in which various reporting processes are incorporated. They were able to show that, excluding under-reporting when present, breeds a case of under estimation of the infectious rate.

In this work, the objective is to develop a model that correctly accounts for cases of under-reporting when given count data. A relative risk map of a given geographical area is plotted with reference to their spatial effects. This is achieved by employing a better method of estimation, i.e. Bayesian method of parameter estimation.

We achieve the above objective by taking some clues from (Waller and Carlin and Xia, & Gelfand, 1997) and assume a different dimension from what (Gamado et. al, 2014) worked on. Under-reporting takes a different turn. In the sense that, under-reporting varies spatially through a probability from region to region and captured by a binomial distribution. In our case, covariates will be excluded with the assumption that, count data varies spatially only.

In section 2, we build the model to cater for under-reported in count data by employing a Negative Binomial distribution instead of Poisson distribution as suggested by (Pararai, Famoye, & Lee, 2010). In Section 3, we derive the parameters using Bayesian method and continue with discussions of results in section 4. This will then be ended by the last but not the least section, dubbed, conclusion and recommendation.

2. Methodology

2.1 Model

Given an independently and identically distributed (i.i.d) trials, a Negative Binomial Distribution can be used to model the number of success before a failure is achieved. In order to correctly model the phenomenon above, we continue to discuss the Negative Binomial (NB) distribution, $Y \sim NB(\pi, r)$. A discrete random variable, Y, can be said to follow a Negative Binomial function if the pdf, $f(.)$, can be written as;

$$f(Y;\pi,r) = \frac{(y+r-1)!}{y!\,(r-1)!}\pi^r(1-\pi)^y,\ 0 \le y \ge r\, and\, r > 0. \tag{1}$$

The mean and variance of the distribution above can be written as $\frac{r(1-\pi)}{\pi}$ and $\frac{r(1-\pi)}{\pi^2}$ respectively. Also, r, an integer-valued parameter, is the number of times ($r = N + y$) that we need to repeat a Bernoulli experiment with success probability, π until N successes are achieved. This makes the dispersion index (DI) as $\frac{1}{\pi}$ where π is the probability of reporting a case or success probability. Also, λ represents the mean. When this happens, $\mathbf{E}(Y) = \lambda = E\pi$. Here, E denotes expected value of the unit under discussion.

Given that $\pi = \left(\frac{r}{r+\lambda}\right)$ and $1-\pi = \left(\frac{\lambda}{\lambda+r}\right)$, the above pdf i.e. Equation 1 transforms into;

$$f(Y;\pi,r) = \frac{(y+r-1)!}{y!\,(r-1)!}\left(\frac{r}{r+\lambda}\right)^r\left(\frac{\lambda}{\lambda+r}\right)^y,\ 0 \le y \ge r\, and\, r > 0. \tag{2}$$

When modelling count data, the assumption of independence for π can be relaxed and made to depend on some covariates. In that case π can be written as;

$$\pi_i = \frac{r}{\lambda+r} = exp\left(u_{1i} + v_{1i} + X'\beta\right). \tag{3}$$

The probability distribution of the spatial effects are;

$$p\left(u_{1i} \mid \mathbf{u}_{-1i}\right) \sim N\left(\frac{\sum_{j\in N(1i)} u_{1j}}{d_{1i}}, d_{1i}^{-1}k_{1i}\right) \tag{4}$$

$$v_{1i} \sim WN\left(0, \sigma_{1i}^2\right) \tag{5}$$

In Equations (3), π_i is the success probability of reporting an event and in our case, it is made to depend solely on structured (u_i) and unstructured spatial effects(v_i) with probability distributions suggested by (Besag, York, Jeremy & Molli, 1991); Ngesa, Achia, & Mwambi, 2014). In Equation (4), N stands for Normal distribution whiles in Equation (5), WN represents White Noise and d_{1i} is the number of neighbouring units.

Count data, mostly come with elements of over-dispersion; this happens when data are collected under non-uniform circumstances. This also happens when the population under study is heterogeneous. In this paper, this is obvious

because of the incorporation of under-reporting. For the data given, the variance is more than seven thousand times the mean, a clear case of over-dispersion.

Also, the marginal expectation of Y_i can be computed as;

$$\mathbf{E}(Y_i|x_i) = E[\mathbf{E}(\pi_i)] = \lambda E. \tag{6}$$

The Variance of Y_i can be written as;

$$\begin{gathered} Var(Y_i|x_i) = \mathbf{E}(Var(Y_i|x_i, \pi_i)) + Var(\mathbf{E}(Y_i|x_i, \pi_{\cdot i})) \\ Var(Y_i|x_i) = \mathbf{E}(Y_i|x_i) + E^2 Var(\pi_i) \end{gathered} \tag{7}$$

$$Var(Y_i|x_i) = \lambda + E^2 Var(\pi_i). \tag{8}$$

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.

To account for over-dispersion and under-reporting, we come up with a joint distribution between the binomial and negative binomial distribution. This can be written as;

$$p(Y_i = y) = \left\{ \sum \frac{(y+r-1)!}{y!(r-1)!} \left(\frac{r}{r+(\pi_u\lambda)}\right)^r \left(\frac{(\pi_u\lambda)}{(\pi_u\lambda)+r}\right)^y \frac{y^*!}{(y^*-y)!y!} (\pi_u\lambda)^y \times (1-(\pi_u\lambda))^{y^*-y}, \right\} \tag{9}$$

$$p(Y_i = y) = \frac{\left(y+\alpha^{-1}-1\right)!}{y!(\alpha^{-1}-1)!} \left(\frac{\alpha^{-1}}{\alpha^{-1}+(\pi_u\lambda)}\right)^{\alpha^{-1}} \left(\frac{(\pi_u\lambda)}{(\pi_u\lambda)+\alpha^{-1}}\right)^y, \; y \geq 0. \tag{10}$$

where π_u is the probability of under-reporting which varies spatially through a binomial probability and α is the inverse of r. This can be written as;

$$logit(\pi_{u_i}) = u_{2i} + v_{2i} \tag{11}$$

with their probability distribution functions as;

$$p(u_{2i} \mid \mathbf{u}_{-2i}) \sim N\left(\frac{\sum_{j\in N(2i)} u_{2j}}{d_{2i}}, d_{2i}^{-1} k_{2i}\right), \tag{12}$$

$$v_{2i} \sim WN\left(0, \sigma_{2i}^2\right). \tag{13}$$

In the equation above, u_{2i} and v_{2i} are the structured and unstructured spatial effects in the under-reporting probability. The average number of observed cases for a period of one year is $\mu = \pi_u\lambda$. This can also be written as;

$$\mu = \pi_u\lambda = \pi_u\left(\alpha^{-1}\left(\frac{1-\pi}{\pi}\right)\right), \tag{14}$$

$$= \alpha^{-1} \frac{\{exp(u_2+v_2) - exp\{(u_1+u_2)+(v_1+v_2)+X'\beta\}\}}{\{exp(u_1+v_1+X'\beta) - exp\{(u_1+u_2)+(v_1+v_2)+X'\beta\}\}}. \tag{15}$$

In this case, our dependent variables are assumed to vary spatially and not on some covariates. In that case, Equation (15) runs into;

$$\mu = \alpha^{-1} \frac{\{exp(u_{2i}+v_{2i}) - exp\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}}{\{exp(u_{1i}+v_{1i}) - exp\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}} \tag{16}$$

With the above in mind, the likelihood function of Equation (10) can be written as;

$$L(y;\pi_u,\alpha)=\prod_{i=0}^{n}\frac{\left(y+\alpha^{-1}-1\right)!}{y!\left(\alpha^{-1}-1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+(\pi_u\lambda)}\right)^{\alpha^{-1}}\left(\frac{(\pi_u\lambda)}{(\pi_u\lambda)+\alpha^{-1}}\right)^{y}, \tag{17}$$

$$\left.\begin{array}{rl}\ln L(y;\pi_u,\alpha) & =\sum_{i=0}^{n}\left\{\ln\left(\left(y+\alpha^{-1}-1\right)!\right)-\ln\left(y!\left(\alpha^{-1}-1\right)!\right)+\alpha\ln\alpha\right\}\\ & +\alpha\ln\left(\alpha^{-1}+(\mu)\right)+y\ln\mu-\ln\left((\mu)+\alpha^{-1}\right).\end{array}\right\} \tag{18}$$

Substituting Equation (16) into Equation (17) gives;

$$\left.\begin{array}{rl}\ln L(y;\pi_u,\alpha)= & \sum_{i=0}^{n}\left\{\ln\left(y+\alpha^{-1}-1\right)!-\ln\left(y!\left(\alpha^{-1}-1\right)!\right)+\alpha\ln\alpha\right\}\\ & +\alpha\ln\left(\alpha^{-1}+\left(\alpha^{-1}\frac{\{exp\,(u_{2i}+v_{2i})-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}}{\{exp\,(u_{1i}+v_1)-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}}\right)\right)\\ & +y\ln\left(\alpha^{-1}\frac{\{exp\,(u_{2i}+v_{2i})-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_2)+X'\beta\}\}}{\{exp\,(u_{1i}+v_{1i}+X'\beta)-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})+X'\beta\}\}}\right)\\ & -\ln\left(\left(\alpha^{-1}\frac{\{exp\,(u_{2i}+v_{2i})-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}}{\{exp\,(u_{1i}+v_{1i})-exp\,\{(u_{1i}+u_{2i})+(v_{1i}+v_{2i})\}\}}\right)+\alpha^{-1}\right).\end{array}\right\} \tag{19}$$

A careful look at Equation (10) shows that, it transforms to Poisson and Geometric when the diversion parameter, $\alpha = 0$ and 1 respectively. Having successfully identified the contributing parameters, we present the candidate models in order of complexity below;

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

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

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

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

2.1.1 Parameter Estimation (Bayesian Approach)

Bayesian method is preferred over the usual frequentist method. This is due to its advantage of suppressing the effects of confounding variables. In this method, a prior distribution $p(\theta)$ is first identified and then likelihood $p(y|\theta)$ is then computed from the data through the popular Maximum Likelihood Estimation. After which the Baye's theorem is invoked to compute the posterior distribution, $p(\theta|y)$. Mathematically, $p(\theta|y)\propto p(\theta)\,p(y|\theta)$ with the constant of proportionality being a marginal distribution written as $\int p(\theta)\,p(y|\theta)\,d\theta$.

In this write-up, the main variables are the structured (u), the unstructured (v) and their unknown variances, $\left(\tau,\sigma^2\right)$.

Applying Bayesian method, The variables are estimated as;

$$p(y\mid y*,\pi,\pi_u,v_{1i}v_{2i},\tau).p(y)=p(y*|,\pi,\pi_u,\lambda)\,p(\pi)\,p(\pi_u)\,p(u_{1i},u_{2i},v_{1i}v_{2i},\tau_{1i},\tau_{2i})\,p(\alpha).$$

The above equation will be used in conjunction with the following prior distributions,

$$p(\alpha_0)\sim N(0,\sigma) \tag{20}$$

Here, α_0 in Equation (20) is assumed to also contain the intercept parameter because they have the same domain of existence, which is the whole number line. It is at the backdrop of this that a Normal distribution of mean, 0 and unknown variance is chosen.

$$p(\pi_u)\sim U(0,1) \tag{21}$$

This makes the posterior to be computed as;

$$\left.\begin{array}{rl} p\left(y \mid y*, \pi, \pi_u, v_{1i}v_{2i}, \tau\right) & \propto \prod_{i=0}^{n} \dfrac{\left(y+\alpha^{-1}-1\right)!}{y!\left(\alpha^{-1}-1\right)!}\left(\dfrac{\alpha^{-1}}{\alpha^{-1}+(\pi_u\lambda)}\right)^{\alpha^{-1}}\left(\dfrac{(\pi_u\lambda)}{(\pi_u\lambda)+\alpha^{-1}}\right)^{y} \\ & \quad \times\begin{pmatrix} y^* \\ y \end{pmatrix}\lambda_i^{y_i}\left(1-\lambda_i\right)^{y_i^*-y_i} \\ & \quad \times exp\left\{-\dfrac{1}{2}\sum_{i=1}^{n}\left(\lambda_i-\psi_{2i}\right)^2\right\}.exp\left\{-\dfrac{1}{2\sigma_{1i}^2}\sum_{i=1}^{n}v_{1i}^2\right\}. \end{array}\right\} \tag{22}$$

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;

$$\left.\begin{array}{rl} p\left(u_{1i} \mid \star\right) & = \dfrac{p\left(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right)}{p\left(u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right)} \\ & \propto p\left(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right) \\ & = p\left(\pi_i \mid u_{1i}, v_{1i}, \tau_1, \tau_2\right) \times p\left(u_{1i}\right) \\ p\left(u_{1i} \mid \star\right) & \propto exp\left\{-\dfrac{1}{2\tau_1}\sum_{i=1}^{n}\left(\pi_i-\psi_{1i}\right)\right\}.exp\left\{-\dfrac{1}{2}\sum_{i}\sum_{j\in N(1j)}\left(u_{1i}-u_{1j}\right)^2\right\} \end{array}\right\}. \tag{23}$$

where ψ_{1i} represents the covariates, $X\beta$ for region 1. From Equation (23), we conclude that the posterior distribution of u_i takes on the form of the marginal distribution of u_i specified in Equation (12).

With $\star$ standing for the conditioning arguments $u_{1i}, u_{2i}, v_{1i}v_{2i}, \tau$, the posterior distribution of the correlated part ,u_{2i}, of the under-reported probability can be computed as;

$$\left.\begin{array}{rl} p\left(u_{2i} \mid \star\right) & = \dfrac{p\left(\pi_u, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right)}{p\left(u_{1i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right)} \\ & \propto p\left(\pi_u, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right) \\ & = p\left(\pi_u \mid u_{2i}, v_{2i}, \tau_1, \tau_2\right) \times p\left(u_{2i}\right) \\ p\left(u_{2i} \mid \star\right) & \propto exp\left\{-\dfrac{1}{2}\sum_{i}\sum_{j\in N(j)}\left(u_{2i}-u_{2j}\right)^2\right\} \end{array}\right\}. \tag{24}$$

We conclude that the posterior distribution of u_{2i} takes on the form of the marginal distribution of u_{2i}.

with that of the uncorrelated part, v_{2i},of the under-reported probability can be computed as;

$$\left.\begin{array}{rl} p\left(v_{2i} \mid \star\right) & = \dfrac{p\left(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right)}{p\left(u_{1i}, u_{2i}, v_{1i}, \tau_1, \tau_2 \mid y\right)} \\ & \propto p\left(\pi_i, u_{1i}, u_{2i}, v_{1i}, v_{2i}, \tau_1, \tau_2 \mid y\right), \\ & = p\left(\lambda_i \mid u_{1i}, v_{2i}, \tau_1, \tau_2\right) p\left(v_{2i}\right), \\ p\left(v_{2i} \mid \star\right) & \propto exp\left\{-\dfrac{1}{2\tau}\sum_{i=1}^{n}\left(\lambda_i-\psi_{2i}\right)^2\right\}.exp\left\{-\dfrac{1}{2\sigma_{1i}^2}\sum_{i=1}^{n}v_{2i}^2\right\}. \end{array}\right\}. \tag{25}$$

We conclude that the posterior distribution of v_{2i} takes on the form of the marginal distribution of v_{2i}.

The conditional posterior distribution of τ can be computed as;

$$\left.\begin{array}{rl} p\left(\tau \mid \star\right) = & p\left(\tau \mid \alpha, \delta\right) p\left(\tau\right) \\ & \propto \left(\tau\right)^{-(\alpha+1)} exp\left(-\frac{\delta}{\tau}\right). \end{array}\right\} \tag{26}$$

2.2 Data and Results

The aim of this study is to estimate the relative risk of diabetes cases with elements of under-reporting. Data describing diabetes cases was 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 and summed by the district. The yearly recordings were used in this work. 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 only for 2014 as the agency did not have all data for later years. There were 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 paper was carried out under the assumption that covariates are not available. WinBugs version 1.4 was used in the implementation (Spiegelhalter , Thomas , Best, & Lunn, 2003) phase. A double chain of Markov chain Monte Carlo (McMC) iterations of 70,000 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 of the remaining 6,000 samples. The decision on convergence was arrived at based on the behaviour of our trace plots and auto-correlation/Time series plots of the McMC output (Gelman et al., 2014). According to (Gelman et al., 2014), when the trace plot for a double chain appears to be crossing each other, then that is an indication for convergence. The posterior means of each model was used in the assessment of their efficiency and then a Deviance Information Criterion (DIC) was generated. The models were compared using the (DIC) as proposed by (Spiegelhalter et al., 2003). The best fitting model is the model with the smallest value of DIC and in this case, Model 4 was used in the analysis.

The data we obtained is such that the variance is far greater than the mean. This could be as a result of over-dispersion. Also, larger variance than mean could also be attributed to the presence of under-reporting as was shown in Equations (6 and 8) . In this work, under-reporting has been catered for by assuming that it varies spatially in all units. Under-reporting varies through a probability captured by a binomial distribution and solely dependent on spatial properties. The overall count was then modelled using a Negative Binomial instead of Poisson. This choice was as a result of the occurrence of elements of over-dispersion. This further translates that, Negative Binomial (NB) is adopted for the estimation of the relative risks.

Apart from the Monte Carlo (MC) Error, the decision of convergence was arrived at by the intertwining nature of Time series Figures (1,3,7,9) and Trace plots (2,4,8,10) plots.

This work was based on the assumption that, both the count data and under-reporting vary spatially for all regions. This spatial properties can be divided into two parts which are correlated, u_1 and uncorrelated parts, v_1. In the case count data, the mean values of the correlated part, u_{1i} fall in the range of $(-2.194, 2.602)$ with most of the 95% credible interval been all positive. This signifies positive relationship with the relative risk. A similar thing happens in the case of the uncorrelated spatial part for the count data, v_{1i} with the mean values falling in the ,margin of $(-1.247, 0.2919)$ with all the 95% credible interval being positive, which is an indication of positive relationship with the relative risk.

The values and the the credible intervals of the generated relative risks are probabilistic in nature, i.e. all below one with the values ranging from $(0.001\ to\ 0.3)$ in the case of the Negative Binomial and that of the Poisson falling in the range of $(0.1\ to\ 50)$. With reference to the variables, u_{1i}, v_{1i}, v_{2i} and u_{2i}, there was a positive relationship between the parameters and the relative risk for all units. With the case of the under-reporting probabilities, (π_u), the correlated part, u_{2i}, falls in the range of $(-0.103, 0.13)$ with all the 95% credible interval being all positive. The uncorrelated part, v_{2i} also falls in the range of $(-1.245, 1.066)$ with all the (95%) being positive. These results projects the importance of incorporating spatial elements when modelling count data.

With reference to Table (1), Model (2) is a better estimate of the count data better than Model (3) because it has a lower DIC. It can also be seen from the table that, the uncorrelated, v_{2i}, part of the probability of under-reporting (π_u) , contributes more to the model than the correlated parts, u_{2i}. This is evident in the DIC between the two models. The fourth model is however the best model with the lowest DIC.

It can also be said that, most of the geographical units under study fall below the (0.05) mark, Figure (1), signifying low risk, with most of these geographical units located in the Southern part of Ghana. The probability of under-reporting, Figure (2) also has most of the geographical units occurring with values greater than 0.1.

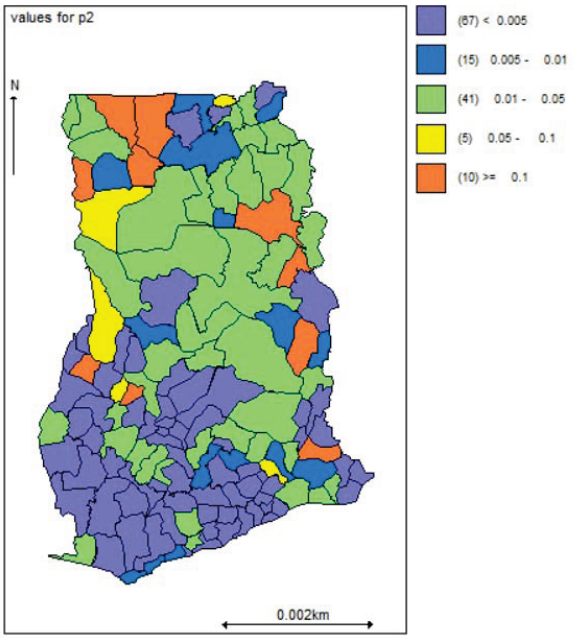

Figure 1. Relative Risk of diabetes cases in Ghana

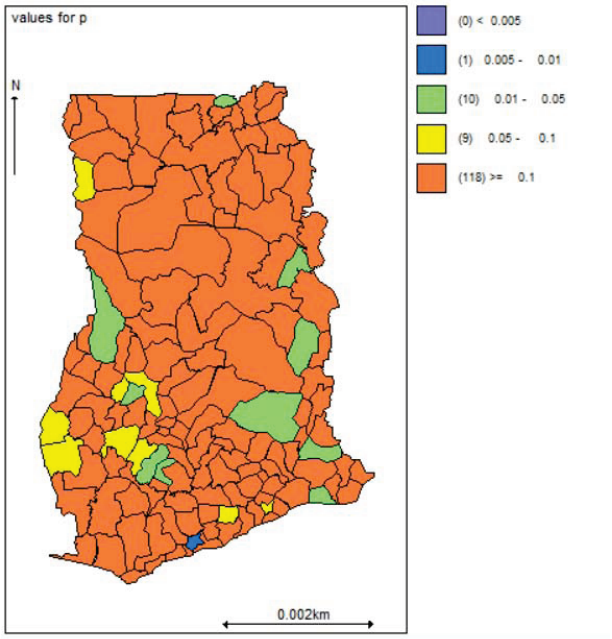

Figure 2. Probability of Underreporting

Table 1. Comparison of Count Models in Ghana

	Model			
	1	2	3	4
α_0	6.571 (6.005, 6.862)	7.22 (6.75, 7.563)	7.336 (6.98, 7.639)	7.261 (6.891, 7.602)
$\sigma^2_{u_1}$	1.506 (0.9257, 2.243)	1.586 (0.9596, 2.354)	1.413 (0.7543, 2.154)	1.523 (0.29012, 2.238)
$\sigma^2_{v_1}$	0.02527 (0.0354, 1.273.)	0.3531 (0.0258, 1.324)	0.3175 (0.02699, 1.062)	0.3414 (0.02836, 1.0822.)
$\sigma^2_{u_2}$	*NA*	*NA*	0.7612 (0.0327, 2.381)	0.4589 (0.03039, 1.822)
$\sigma^2_{v_2}$	*NA*	1.62 (0.05458, 2.761)	*NA*	1.544 (0.05795, 2.738)
pD	43	−499.408	−240	−686.631
DIC	2141.2	1265	1757.520	1020.500

5. Conclusion and Recommendation

The addition of spatial variables to the model produces very good fit for the data. This we did by introducing structured and unstructured spatial elements in both the count data and an under-reporting probability. The spatially structured random effects was captured by the usual Conditional Auto-regressive (CAR) model proposed by (Besag et al., 1991; Ngesa et al., 2014). From the results obtained from Figure (1), it could be seen that most of the high risk areas were identified to be in the southern regions of Ghana. This outcome can be attributed to the eating and behaviour pattern of the inhabitants there. (Darkwa, 2011) pin-pointed regions from the southern part of Ghana as regions with high risk of contracting diabetes; with a risk higher that the world average. Interestingly, inhabitants in the south have their stable foods prepared from maize, rice, cassava and yam. These crops are proven to fuel diabetes in humans (Darkwa, 2011).

The low risk areas on the other hand are known with eating meals prepared from crops like millet, sorghum and guinea corn. These are foods known with low carbohydrates. Also, people in the north are known to trek long miles to their farms serving as exercise for them. Exercising is a good therapy for diabetic patients (Darkwa, 2011; Danquah et al., 2012).

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 although it can be factored in future works.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 exist in each geographical setting.

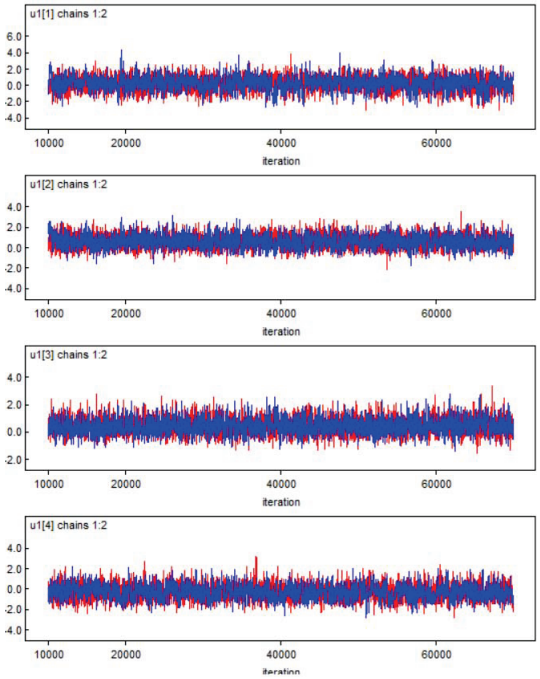

Figure 3. Time series plot of u_1

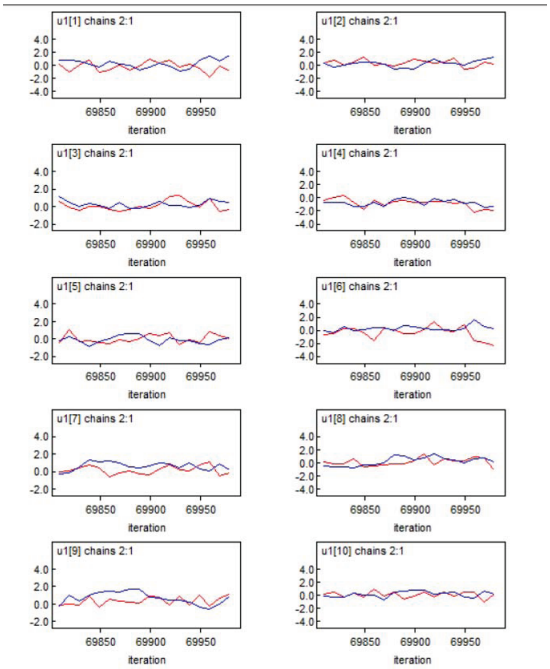

Figure 4. Trace plot of u_1

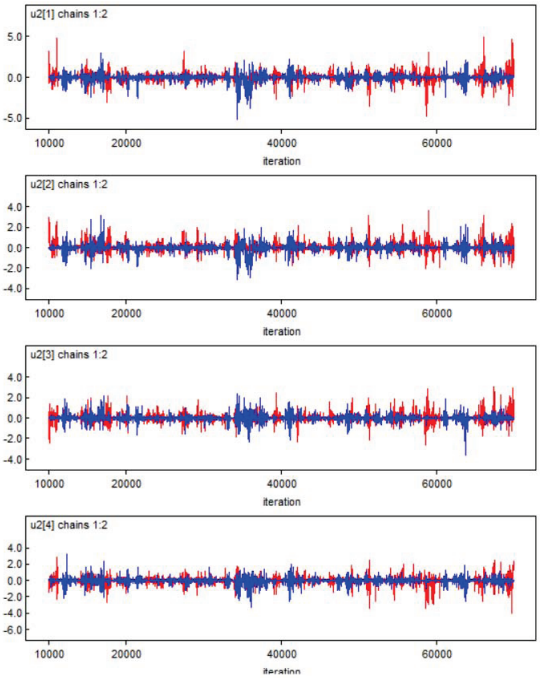

Figure 5. Time series plot of u_2

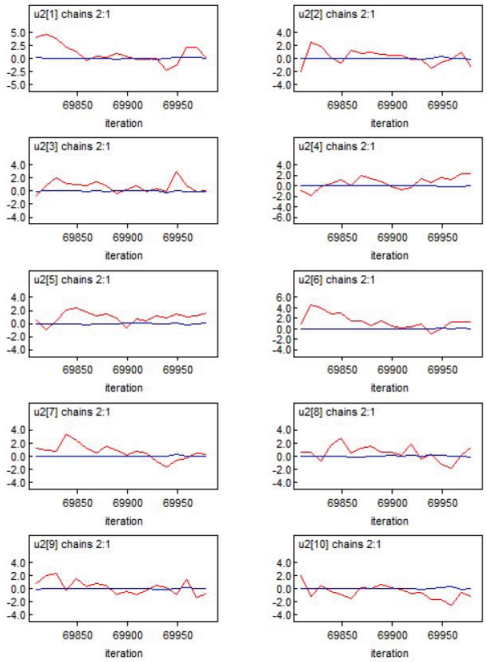

Figure 6. Trace plot of u_2

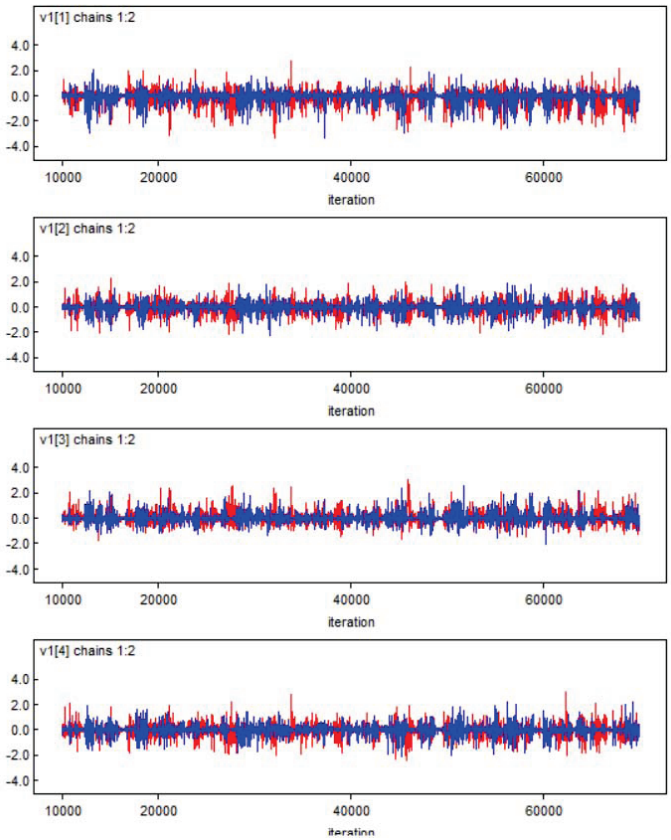

Figure 7. Time series plot of v_1

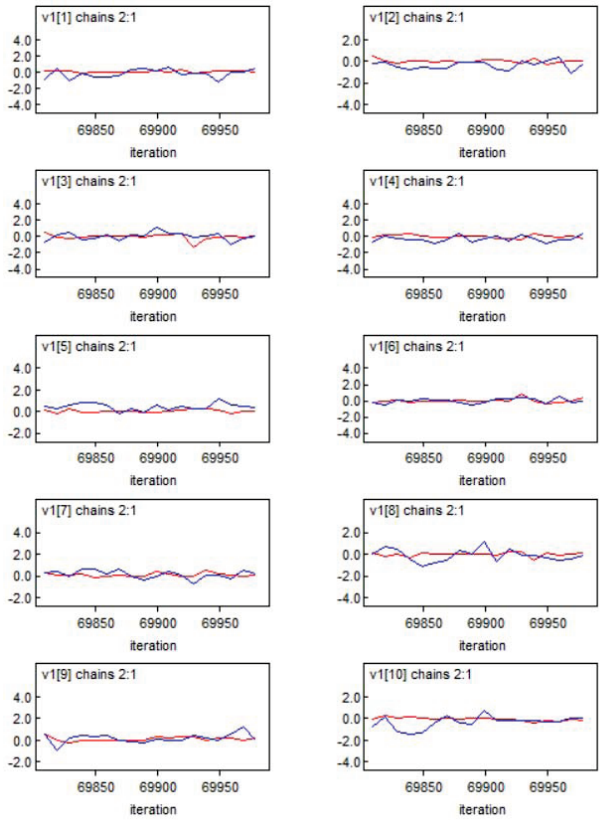

Figure 8. Trace plot of v_1

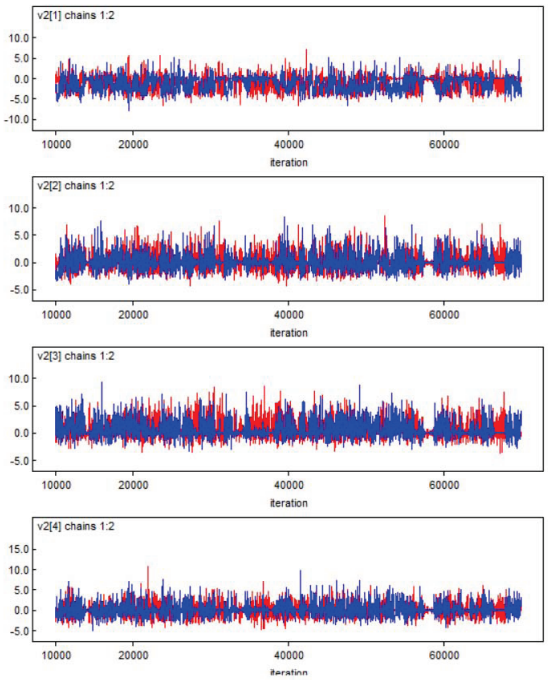

Figure 9. Time series plot of v_2

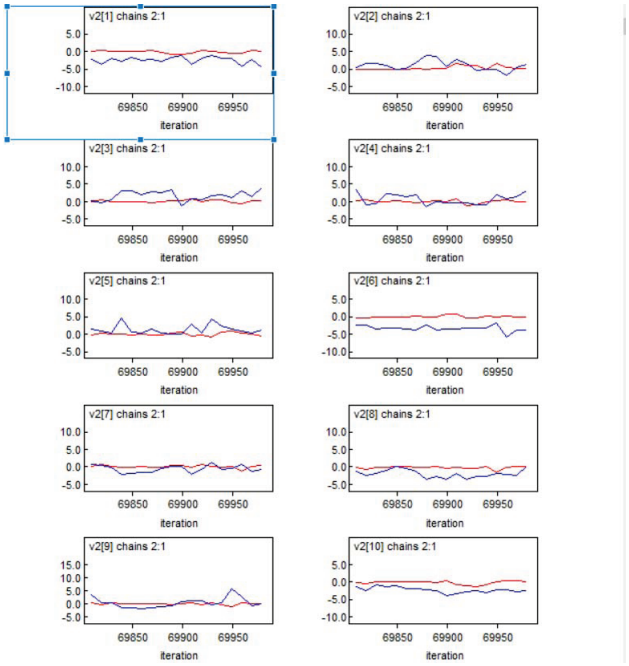

Figure 10. Trace plot of v_2

Acknowledgements

I wish to extend my thanks giving to the Almighty God. Also, to my Uncle Mr. Appiah and his wife.

References

Besag, J., York, Jeremy, & Molli, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics, 43*(1), 1–20. https://doi.org/10.1007/BF00116466

Danquah, I., Bedu-Addo, ..., & Dietz, E. (2012). Diabetes mellitus type 2 in urban Ghana: characteristics and associated factors. *BMC Public Health, 12*(1), 1. https://doi.org/10.1186/1471-2458-12-210

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.

Gamado, K. M., Streftaris, G., & Zachary, S. (2014), Modelling under-reporting in epidemics. *Journal of mathematical biology, 69*(3), 737–765. https://doi.org/10.1007/s00285-013-0717-z

Gelman, A., Carlin, J. B., ..., & Donald, B. (2014). *Bayesian data analysis, 2*(Chapman and Hal/ CRC Boca Raton, FL, USA.)

Moore, D. A., & Carpenter, T. E. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. *Epidemiologic reviews, 21*(2), 143–161. https://doi.org/10.1093/oxfordjournals.epirev.a017993

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

Pararai, M., Famoye, F., & Lee, C. (2010). Generalized Poisson-Poisson Mixture Model for Misreported Counts with an Application to Smoking Data. *Journal of Data Science, 8*(4), 607–617.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual*.

Ugarte, M. D., Ibez, B., & Militino, A. F. (2006). Modelling risks in disease mapping. *Statistical Methods in Medical Research, 15*(1), 21–35. https://doi.org/10.1191/0962280206sm424oa

Waller, L. A., Carlin, B. P., ..., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association, 92*(438), 607–617. https://doi.org/10.1080/01621459.1997.10474012

Winkelmann, R. (1996), *Markov chain Monte Carlo analysis of underreported count data with an application to worker absenteeism*. Empirical Economics. https://doi.org/10.1007/BF01180702