Childhood malaria in the Gambia: A case-study using generalized linear model with Bayesian methods

Xuzhi Wang, Tana Gegen, Mengqiu Zhu and Jizhou Kang

December 22, 2017

1 Introduction

1.1 Background

Malaria is an environmental disease, it's a major public health issue in much of the developing world. As mentioned in Thomson et al. (1999) [6], understanding how different kinds of factors influence the prevalence of malaria is important. Besides, what we care about is not only the influential factors but the mapping of its risk geographically to improve the targeting of scarce resources for public health interventions, since it allows interventions to be adapted to specific sites. This is essential for the effective control of the disease.

Our objective is to find out the factors that could influence and how they influence the prevalence of malaria in a sample of villages in Gambia and describe the model's spatial variation in the prevalence of malaria to provide an explanation of the residual extrabinomial variation. So we can assess whether the extrabinomial variation is spatially structured.

The dataset we use is from Ribeiro and Diggle (2016) R package 'geoR' [5]. It was collected during the second year (1992) of a study designed to measure the effectiveness of the National Impregnated Bednet Programme. It describes Malaria prevalence in children recorded at villages in The Gambia, Africa.

The original data set is based on individual level, however, we transformed the data into village level. Because it's more important to find out how malaria spreads across villages in an area so necessary interventions can be applied to certain locations, which we believe is more important than studying its effect on individuals.

Diggle et al. (2002) [1] developed a spatial generalized linear mixed model to describe the variation in the prevalence of malaria among a sample of village resident children in the Gambia. That model included individual level covariates, village level covariates as well as separate components for residual spatial and non-spatial extrabinomial variation. The final model takes the following form:

$$log\left\{p_{ij}/(1-p_{ij})\right\} = \alpha + \beta^{'} z_{ij} + S(x_i) \tag{1}$$

Our model is based on Diggle's result but since we are using village level data, it differs slightly in the model assumption of generalized linear model and noise term. Our models are discussed in detail in section 1.2.

1.2 Methods

Our model approach simply follows the procedure of building generalized linear model. Firstly, we assume the number of children having malaria in each village follows a binomial distribution: Binomial(n_i, p_i), where n_i is the total number of samples in a village and p_i is the mean parameter that will be regressed by covariates.

Then, to model p_i, we adapted simple linear regression model but assume different noise terms,

1. The village level noise term is independent, identically distributed normal random variable with mean zero and unknown variance.
2. The village level noise term is spatial process term $S(x)$ plus independent noise.
3. The village level noise term is just an unobserved spatial process.

Detailed description of each model will be presented in the following sections, and we will compare them to draw our conclusion.

When we develop all the models above, we apply Bayesian methods in all of them. The reasons are as follows: by Bayesian methods, we can get the posterior distribution of all the parameters that we are concerned about instead of getting the point estimation of all the parameters, which helps us deal with parameter uncertainty properly. Secondly, we can make use of all the information available and take into account our initial beliefs about our model. Thirdly, we are going to use Gaussian process to measure the spatial effect, in which the maximum likelihood estimator of range parameter is hard to get. Using a Bayesian approach and sampling from the posterior distribution is our only choice.

1.3 Data Description and Choice of Features

The features we choose are described as in Table 1.

Features	Description
green	greenness of surrounding vegetation as derived from satellite
PHC	whether there is a primary health care system in a village
netuse	the number of children who use a bed net in a village
treated	the number of children who use a bed net that is treated
region	central, eastern and western regions (dummy variable)
(x,y)	coordinates of the villages

Table 1: Choice of features

We choose the features above because green is an index about the environmental condition and the prevalence of malaria is an environmental disease; PHC represents whether a village can respond to malaria effectively and prevent its prevalence in time; mosquitoes transmit this highly infectious disease so the use of bed net as well as treated bed net is an essential factor. There are 3 regions: central, eastern, and western regions. The coordinates of the villages are also taken into consideration. The response variable is the number of children with malaria in a village.

1.4 Visualization of Features

In this section, we will have a general idea about how the features that we choose can influence the prevalence of malaria by some plots. First we plot the features onto the maps.

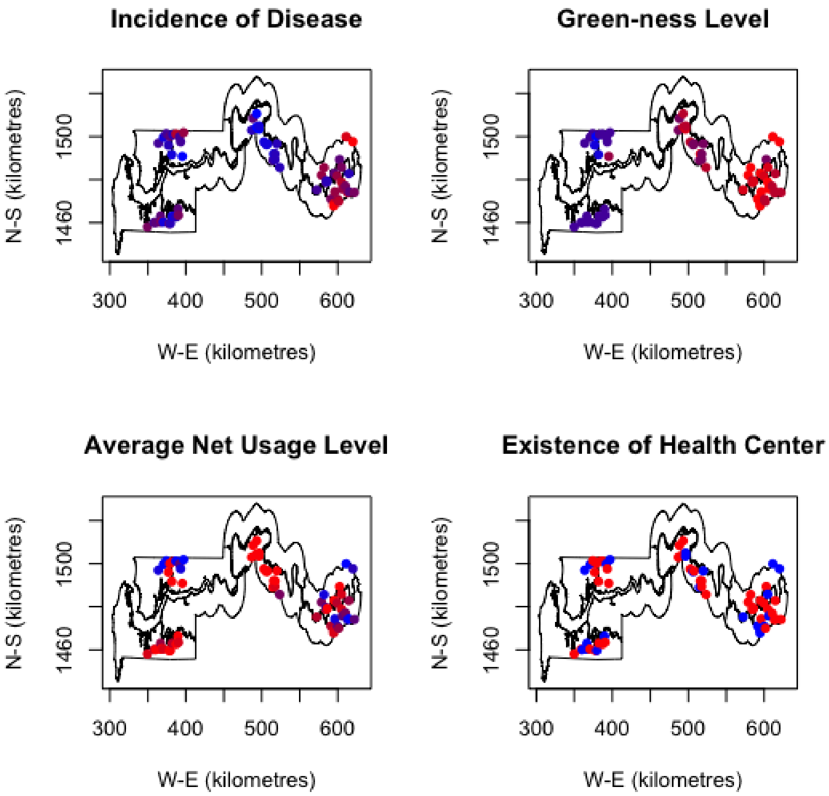

Figure 1: Maps of prevalence of malaria and features

As is shown in Figure 1, it shows four maps of incidence of malaria, green-ness level, average net usage level and existence of health center. The maps above show that:

1. The village we choose are clustered in five areas, three regions: central, eastern and western;

2. Features are clustered according to spatial locations. The values of incidence of disease and the values of other features are close to each other at locations close to each other.

3. The correlation of prevalence of malaria and different features can be approximated roughly, for example, in the central area where the incidence of disease is relatively low, the corresponding net use is relatively high.

Next several figures show how different features can influence the prevalence of malaria respectively.

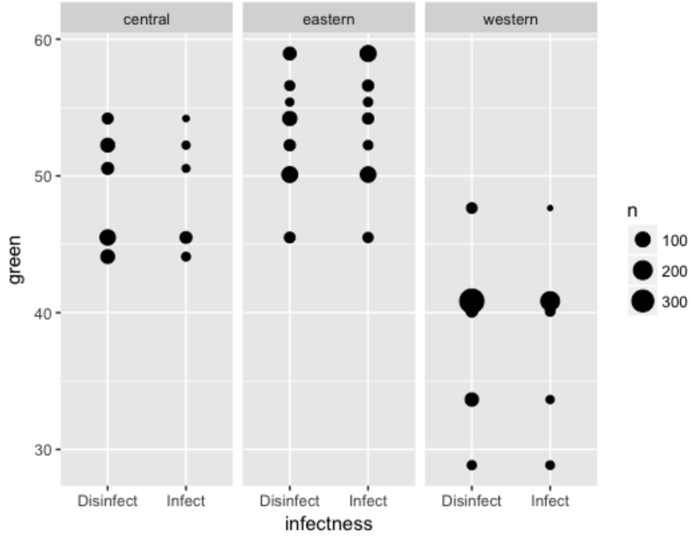

Figure 2: Greenness

Note that in Figure 1, the villages are clustered in 5 areas and those five areas can be divided into 3 regions, central, eastern and western. The feature of green-ness level is also clustered in those 3 regions, the average level in the eastern region is high, while it's middle in the central region and low in the western region. So Figure 2 shows green-ness level in different regions, however, there is no clear trend between the incidence of disease and the green-ness level.

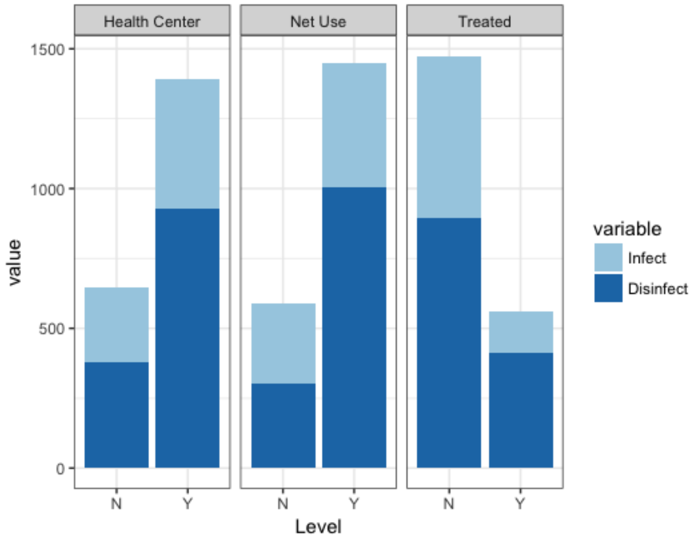

Figure 3: Features and incidence

In Figure 3, we show how PHC, netuse and treated these 3 variable influence the prevalence of malaria. Take PHC for example, we can see that the ratio of disinfected people to infected people is obviously larger in the villages with PHC than that in the village without PHC. As for the other 2 variables, the influence is similar. So if a villages has a health center and more people use bed net and treated bed net, the incidence of malaria will be significantly reduced.

Figure 4 shows the correlation among these features. Basically all the values of correlation are almost about or below 0.25, indicating there is no multicollinearity. So for this dataset, we can use linear models with these features.

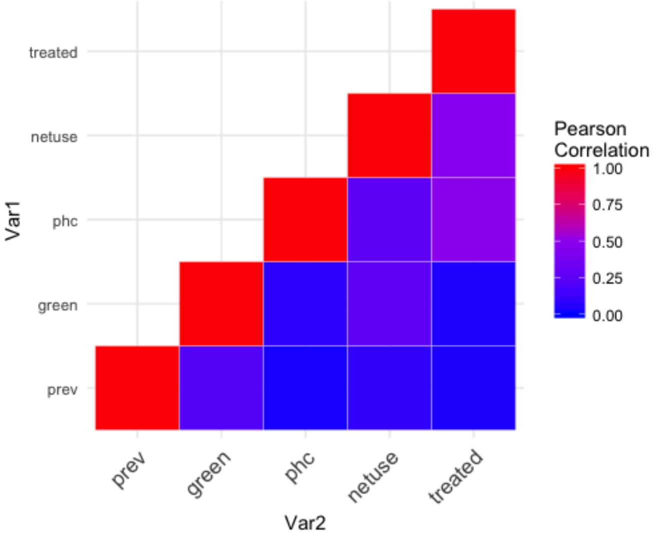

Figure 4: Correlations

The rest of this report will be organized as follows: Section 2 through Section 4 will present our three models together with parameter estimation and diagnose. In Section 5 we will compare the result of three models and draw our final conclusion.

2 Model with random noise

2.1 Model

Following the procedure of building generalized linear model and mixing effect model as mentioned in Hoff's book [3], for the i-th village, the response y_i is the number of children with malaria, n_i is the number of children, x_i is the vector of covariates. We propose a two-level logistic regression model:

$$\begin{aligned} &y_i \sim \text{Binomial}(n_i, p_i) \\ &z_i = \log \frac{p_i}{1-p_i} \\ &z_i = x_i^T \beta + \epsilon_i \\ &\epsilon_i \sim N(0, \sigma^2) \\ &\pi(\beta, \sigma^2) \propto 1/\sigma^2 \end{aligned} \tag{2}$$

We assume that y_i follows a binomial distribution with probability p_i for each child to catch malaria in the i-th village. The link function, z_i, is modeled as the response variable of a linear model with random noise ϵ_i, which follows a normal distribution. In the end, for the joint distribution of β and σ^2, we propose an non-informative prior distribution with density proportional to $1/\sigma^2$.

The joint posterior of parameters is as follows:

$$p(\boldsymbol{z}, \beta, \sigma^2 \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \beta, \sigma^2) \pi(\beta, \sigma^2) \tag{3}$$

Here, $\boldsymbol{y} = (y_1, y_2, ..., y_M)^T$ and $\boldsymbol{z} = (z_1, z_2, ..., z_M)^T$, where M is the number of villlages.

Since ϵ_i's are independent, we can write the likelihood function as:

$$\begin{aligned} L(\boldsymbol{z}, \beta, \sigma^2) \propto \prod_{i=1}^{65} & \exp\{z_i y_i - n_i \log[1 + \exp(z_i)]\} \times \\ & (\sigma^2)^{-1/2} \exp[-\frac{1}{2\sigma^2}(z_i - x_i^T \beta)^2] \end{aligned} \tag{4}$$

We apply MCMC algorithms to sample from posterior distributions of the parameters. To be more specific, we sample β and σ^2 using full conditionals Gibbs sampler and z_i's using Metropolis algorithm.

Full conditional distribution of β and σ^2:

$$\beta \mid z, \sigma^2 \sim \text{MVN}[(X^T X)^{-1}(X^T z), (X^T X)^{-1} \sigma^2] \tag{5}$$

$$\sigma^2 \sim \text{Inv-Gamma}[M/2, \text{SSR}(\beta)/2] \tag{6}$$

Full conditional probability of z_i's:

$$p(z_i \mid y_i, \beta, \sigma^2) \propto \exp\{z_i y_i - n_i \log[1 + \text{exp}(z_i)] - \frac{1}{2\sigma^2}(z_i - x_i^T \beta)^2\} \tag{7}$$

2.2 Parameter Estimation

We ran the MCMC for 50,000 times with 5000 times in the burning period, thus obtaining a sample of 45,000 values from the posterior distribution of each β parameters. From the trace plots in Figure 5, we may conclude that the Markov chains for all 7 β parameters have converged after the burning period. The distribution plots in Figure 6 all show norm shapes as expected.

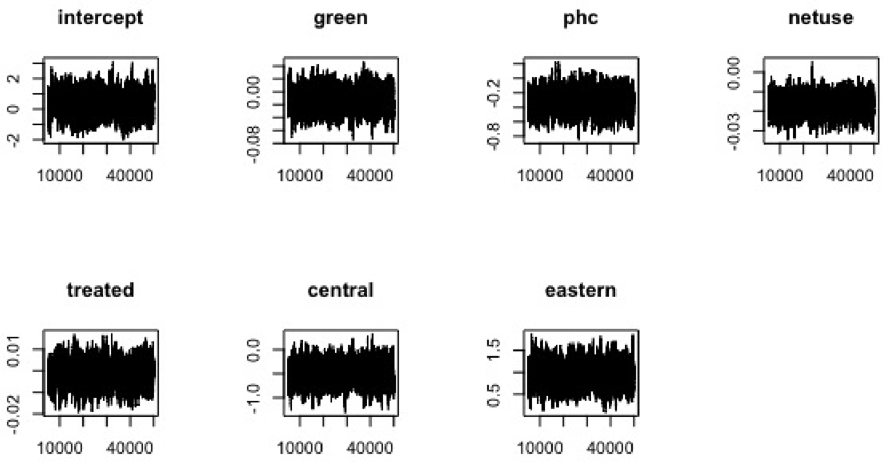

Figure 5: Trace plot: Non spatial model

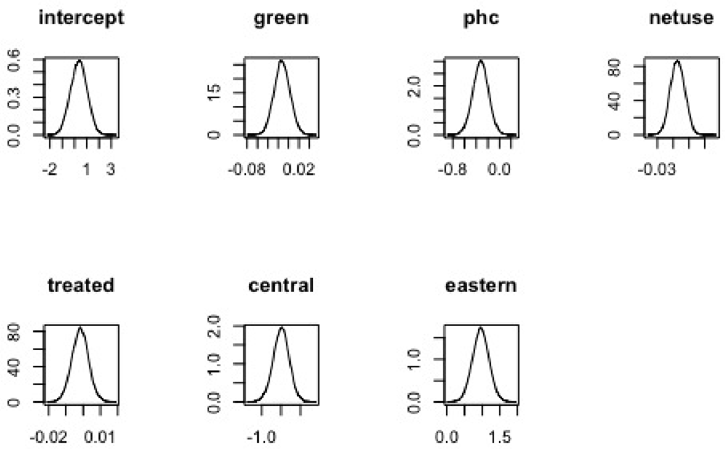

Figure 6: Posterior distribution plot: Non spatial model

	2.5% quantile	97.5% quantile	Mean	Median
intercept	-0.94860	1.64172	0.35072	0.35681
green	-0.04192	0.01746	-0.01250	-0.01263
phc	-0.58766	-0.05826	-0.32301	-0.32189
netuse	-0.02573	-0.00834	-0.01709	-0.01714
treated	-0.01134	0.00795	-0.00174	-0.00172
central	-0.92170	-0.09241	-0.49720	-0.49363
eastern	0.49899	1.42551	0.95853	0.95872

Table 2: Posterior Distribution table: Non spatial model

Table (2) summaries the the 95% confidence interval, mean, and median of the posterior distribution for all β parameters. As indicated in the table, parameters of presence of health centers, bed net use, central area, and eastern area are the ones whose confidence intervals do not contain 0. Hence, these 4 features are significant from Bayesian perspective. Since the confidence intervals of ”phc”, ”netuse”, and ”central” only contain negative values, we may conclude that a village will have a lower malaria prevalence rate if it gets a new health center and has more children using bed nets, and on average, a village located in the central area has a lower prevalence rate than villages located in the other areas. Also, since the confidence interval of ”eastern” only contains positive values, we can say that, on average, a village located in the eastern area has a higher prevalence rate than villages located in the other areas. The conclusions are consistent with the patterns discovered on maps in Figure 1 and bar charts in Figure 3.

2.3 Diagnostics

We estimated the parameters using their posterior means and then use the estimated parameters to calculate the estimated prevalence rate for each village: $\hat{p}_i = \exp(x_i^T \hat{\beta}_i)/(1 + \exp(x_i^T \hat{\beta}_i))$. We then calculated estimated residuals of prevalence rate for each village as $(y_i/n_i - \hat{p}_i)$. Figure 7 is a plot of the estimated residuals against their corresponding estimated prevalence rates. From the plot, the residuals are roughly randomly distributed against the fitted values.

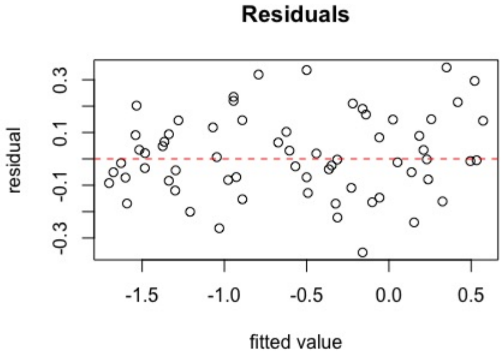

Figure 7: Residuals: Non spatial model

However, when we plotted the residuals on the map, as shown in Figure 8, we discovered that, in each clusters, villages located close to each other tend to have the same level of residuals and villages located far from each other tend to have different levels of residuals.

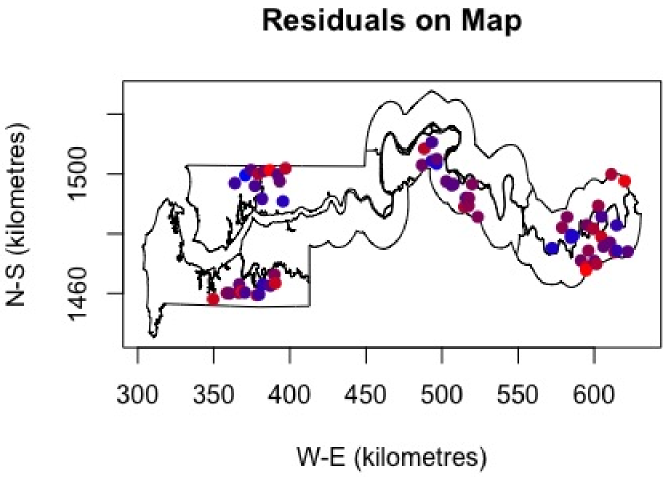

Figure 8: Residuals on map: Non spatial model

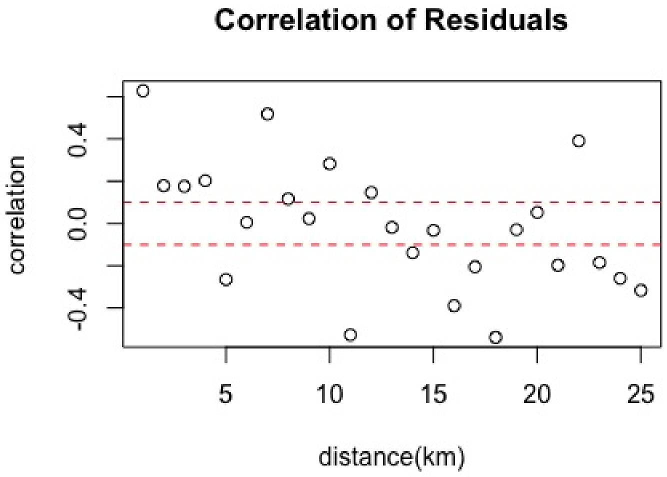

Figure 9: Correlation of Residuals

Therefore, we further calculated the correlation between residuals of pairs of village as a function of the distance between each pair of villages. As shown in Figure 9, villages that are relatively close to each other, for example, within 10 kilometers, have highly positively correlated residuals, while villages relatively far from each other, for example, within 15 to 20 kilometers, have highly negatively correlated residuals. Therefore, it is reasonable to introduce spatial structure into the linear model for the link function z_i's.

3 Model with random noise and spatial factor

3.1 Model

To account for the correlation between residuals of pairs of villages, as suggested in [4], we introduce a spatial factor, S_i as a Gaussian process, into the random noise model:

$$\begin{aligned}
&y_i \sim \text{Binomial}(n_i, p_i)\\
&z_i = \log \frac{p_i}{1-p_i}\\
&z_i = x_i^T\beta + S_i + \epsilon_i\\
&\epsilon_i \sim N(0, \sigma_0^2)\\
&\pi(\beta, \sigma^2) \propto 1/\sigma^2\\
&S_i \sim GaSP(0, \sigma^2 c(\cdot, \cdot))
\end{aligned} \tag{8}$$

We model the spatial factor S_i as a Gaussian Process and choose the covariance function $c(\cdot, \cdot)$ to be Mat$\acute{e}$rn family family with roughness parameter $\frac{5}{2}$. For the hyperparameters in Gaussian process, we use the joint robust prior:

$$\pi^{JR}(\psi_1, \psi_2, \eta) = c(\sum_{l=1}^{p} C_l\psi_l + \eta)^a \exp(-b(\sum_{l=1}^{p} C_l\psi_l + \eta)) \tag{9}$$

Here $\eta := \frac{\sigma_0^2}{\sigma^2}$,and ψ_1, ψ_2 are inverse range parameters.

The joint robust prior, as discussed in [2], is a proper prior which approximate the tail rate of reference prior. It's still a non-informative prior but has benefit such as known normalizing constant and posterior propriety.

Follows our notation, the likelihood function is,

$$\begin{aligned}
L(\mathbf{z}, \beta, \sigma^2) \propto \prod_{i=1}^{65} & exp\{z_i y_i - n_i \log[1 + \text{exp}(z_i)]\}\times\\
& (1/\sigma^2)^{1/2}|R|^{-\frac{1}{2}}\text{exp}[-\frac{(z - X\beta)^T R^{-1}(z - X\beta)}{2\sigma^2}]
\end{aligned} \tag{10}$$

where the only difference with previous model is we have correlation matrix R instead of identity matrix, where

$$R = \sigma^2\mathbf{R}_S + \eta\mathbf{I}, \quad \mathbf{R}_S^{(i,j)} = c(\mathbf{x}_i, \mathbf{x}_j) \tag{11}$$

Similar to the first model, we update β and σ^2 using Gibbs sampling with full conditional distribution

$$\beta^* \sim \text{MVN}[(X^T R^{-1} X)^{-1} X^T R^{-1} z, (X^T R^{-1} X)^{-1}\sigma^2] \tag{12}$$

$$\sigma^{2^*} \sim \text{Inv-Gamma}[M/2, \text{SSR}(\beta)/2] \tag{13}$$

and all the other parameters using Metropolis algorithm.

3.2 Parameter Estimation

After the burning period, the Markov chains of all parameters have converged, and their posterior distribution plots, shown in 10, all have normal shapes as expected.

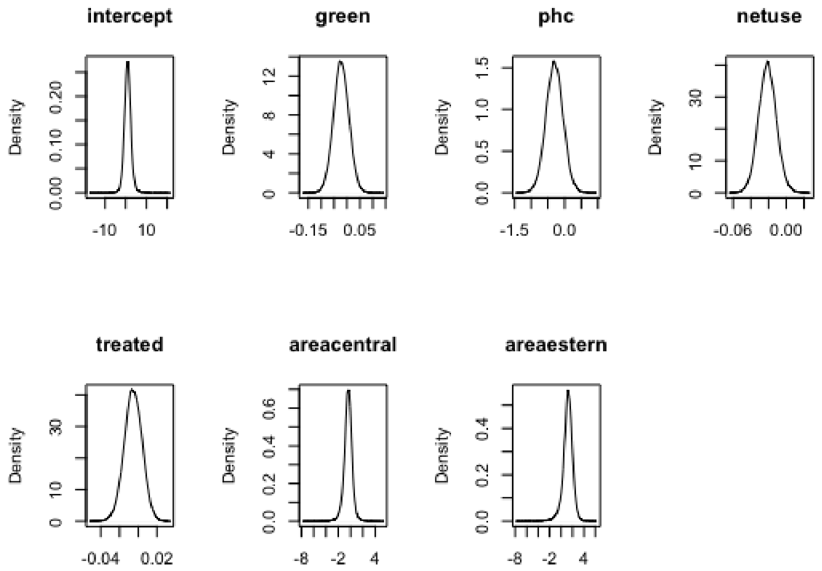

Figure 10: Posterior distribution plot: Spatial model

In Table 3, parameters colored in red are significant in the spatial model, while parameters colored in blue are significant in the non-spatial model but nonsignificant in the spatial model. We can tell that the use of bed net and whether or not the village is located in the eastern area are not relevant in predicting the probability of a child to catch a malaria in the non-spatial model.

	2.5% quantile	97.5% quantile	Mean	Median
intercept	-2.17	4.33	0.98	0.96
green	-0.08	0.04	-0.02	-0.02
phc	-0.81	-0.20	-0.30	-0.30
netuse	-0.04	-0.00	-0.02	-0.02
treated	-0.02	0.01	-0.00	-0.00
central	-1.82	-0.85	-0.41	-0.39
eastern	-0.83	2.72	1.18	1.24

Table 3: Posterior Distribution table: Spatial model

3.3 Diagnostics

Similar to the non-spatial model, we plotted the residuals against fitted probability values and they show no apparent correlation according to Figure 11.

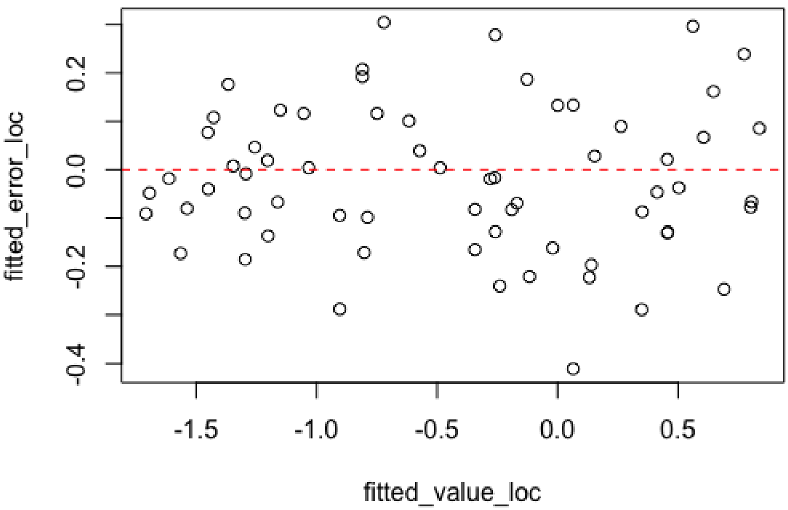

Figure 11: Residuals: Spatial model

4 Model with only spatial factor

4.1 Model

In this case, we remove the random noise term and denote as the reduced model. The likelihood function is as follow and we use the same method to update all the parameters.

$$\begin{aligned} &y_i \sim \text{Binomial}(n_i, p_i) \\ &z_i = \log \frac{p_i}{1-p_i} \\ &z_i = x_i^T \beta + S_i \\ &\pi(\beta, \sigma^2) \propto 1/\sigma^2 \\ &S_i \sim GaSP(0, \sigma^2 c(\cdot, \cdot)) \end{aligned} \tag{14}$$

4.2 Parameter Estimation

We ran the MCMC for 30,000 times with 3000 times in the burning period, thus obtaining a sample of 27,000 values from the posterior distribution of each β parameters. From the trace plots, we conclude that the Markov chains for all 7 β parameters have converged after the burning period. Posterior distributions for each of the corresponding regression parameters are shown in Figure 12.

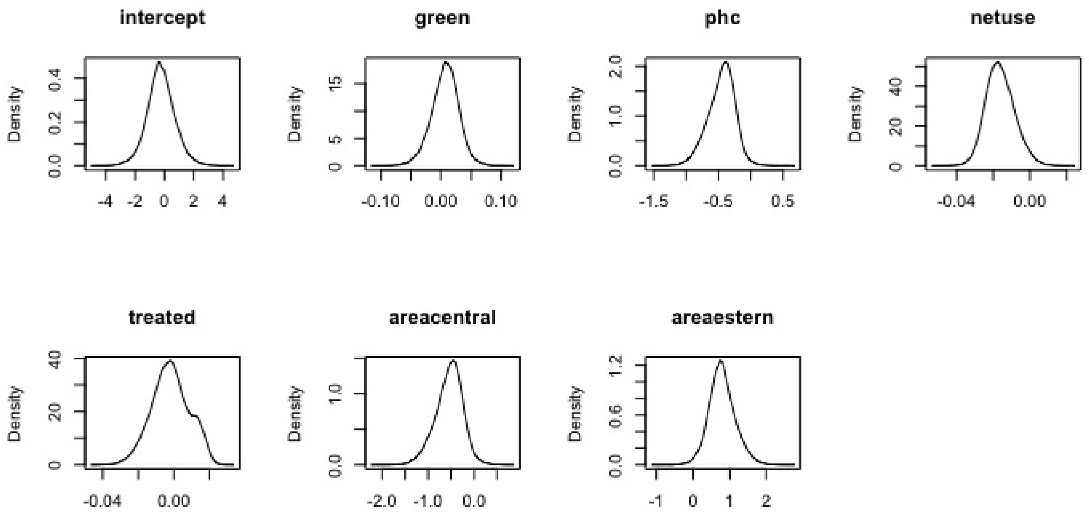

Figure 12: Posterior distribution plot: Spatial model without noise

Table 4 shows the posterior mean, median and 95% credible interval for each of the parameters in model 14. With regard to the regression parameters, it shows that the prevalence of malaria decreases with the presence of health center and village in central part, and that children living in eastern village are more likely to have malaria. Also, the inclusion in the greenness of the surrounding vegetation, bed net use, and treated term do not appear markedly to affect the prevalence of malaria, as in each case the 95% posterior interval comfortably straddles zero. Whether considering spatial effect or not, these three variables do not help us to interpret and predict the probability of each child to catch malaria in each village.

	2.5% quantile	97.5% quantile	Mean	Median
intercept	-2.11915	1.74139	-0.23214	-0.26201
green	-0.04090	0.04941	0.00675	0.00773
phc	-0.90624	-0.45725	-0.32301	-0.43618
netuse	-0.03021	-0.00041	-0.01611	-0.01662
treated	-0.02283	0.01803	-0.00186	-0.00204
central	-1.18688	-0.01260	-0.54503	-0.51898
eastern	0.11395	1.56425	0.79437	0.77419

Table 4: Posterior Distribution table: Spatial model

4.3 Diagnostics

Similarly, Figure 13 plots residual r_i against fitted values and shows the desired absence of any obvious relationship, indicating an adequate fit.

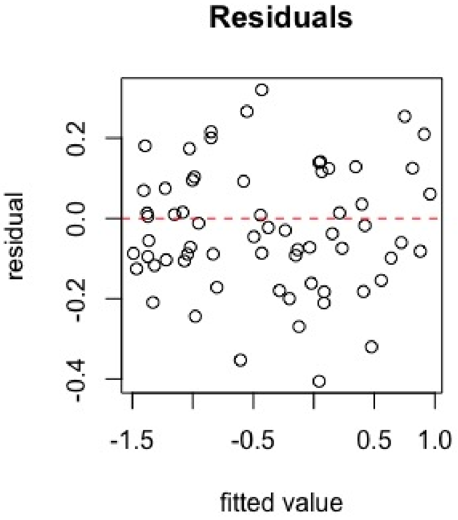

Figure 13: Residuals: Spatial model without noise

5 Model Comparison and Concluding Remarks

5.1 Model Comparison

It is not particularly satisfactory to analyze a single model, but more appropriate to analyze several competing models and compare the results. Thus, we build three models with or without spatial effect and random noise. We compare these models by evaluating their good of fitness to the data. For binomial regression model, when testing goodness of fit, we use the fact that, provided the model under consideration is correct the scaled deviance and the Pearson's X^2 statistics are both asymptotically χ^2 distributed with degrees of freedom equal to the the difference between the number of observation and number of parameters. A response y_i(assumed to be binomial with number of trials m_i) with fitted values $\hat{y}_i$ will have Pearson residuals $r_i^{(P)}$ and deviance residuals $r_i^{(D)}$ given by

$$\begin{aligned} r_i^{(P)} &= \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i/m_i)}} \\ r_i^{(D)} &= sign(y_i - \hat{y}_i)\{2y_i log\frac{y_i}{\hat{y}_i} + 2(m_i - y_i)log(\frac{m_i - y_i}{m_i - \hat{y}_i})\}^{1/2} \end{aligned} \tag{15}$$

corresponding to the Pearson's X^2 statistic and deviance are

$$X^2 = \sum_i \frac{y_i - m_i\hat{\pi}_i}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \tag{16}$$

where $\hat{y}_i = m_i\hat{\pi}_i$ and

$$D(y_i, \hat{y}_i) = 2\sum_i (y_i log(\frac{y_i}{\hat{y}_i}) + (m_i - y_i)log(\frac{(m_i - y_i)}{(m_i - \hat{y}_i)})) \tag{17}$$

We calculate the deviance and Pearson's statistics for the three models and find that the values all exceed 200, far greater than their degrees of freedom. With one possible reason for the large deviance is that, as our data are related to village level with location information, which may lead to overdispersion due to clustering. Therefore, we need to model overdispersion by estimating the dispersion parameter ϕ as follow:

$$\phi = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} = \frac{X^2}{n-p} \tag{18}$$

The estimated dispersion parameters for three models are 3.59, 3.83 and 3.77, respectively. The effect of the parameter ϕ is to increase the estimated variance of the estimated parameters $\hat{\beta}$ but it has no effect on $\hat{\beta}$ itself.

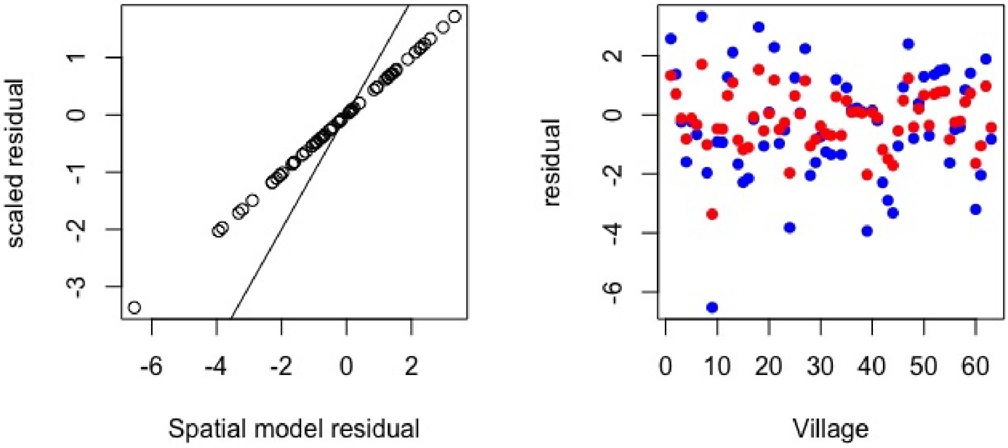

Figure 14: Spatial model residual with and without scale

The dispersion parameter ϕ for model with spatial factor is 3.77, in this case, the red point with scaled residual is smaller than the blue point without scaling. Then, we can use the scaled deviance to perform hypothesis test as follow.

H_0 : Model M_i fits vs H_A : Model M_i does not fit.

Correspondingly, we get the scaled deviance are 58.63, 58.86 and 59.52 for three models. Compared with the degrees of freedom 58 or χ^2_{58}, the scaled deviance are small enough and we fail to reject the null hypothesis, i.e. all three models fit the data.

Since all three models fit the data relatively well, we prefer the third model with only spatial process as the noise term. The reasons are as follows. Firstly, based on our explanatory data analysis, adding spatial effective is quite reasonable and also essential in this case study. Therefore, we prefer the second model and the third model rather than the first one. In addition, in terms of the second and third model, we choose the third one because, based on the test for good of fit, it is not necessary to add the independent noise term and actually adding that term will make the model too flexible which causes interpretation issue.

5.2 Conclusion

In this final project we studied the prevalence of malaria in Gambia. We followed the procedure taught in lecture to build a hierachical Bayesian model to measure the effect of features such as: bednet usage, treatness of bednet, greenness level, and existence of health care center, to the prevalence of malaria. We first assume the number of children having malaria in each village follows a binomial distribution, and we use linear model with different noise term to model the mean parameter after transformed by logistic link function. The three different noise term we considered are: independent noise, spatial process modeled by Gaussian process and the combination of them. We use the MCMC approach to sample from the posterior distribution and based on the corresponding empirical distribution to draw our conclusion. Based on the deviance measure, all three model fits the data well, and we choose the model with only spatial effect as the noise term because spatial correlation is important in this case as well as this model have no identifiable issue. From the parameter estimation result of that model, we conclude that existence of health center and the usage of bednet have significant influence on the prevalence of malaria, and the negative sign means use bed net and having health center will help prevent malaria. The coefficients for categorical variable of location suggests that eastern area has highest frequency of malaria. Therefore, the government should spend more money in distributing bednet and building health care center, as well as focusing more on eastern area to prevent malaria.

References

[1] Peter Diggle, Rana Moyeed, Barry Rowlingson, and Madeleine Thomson. Childhood malaria in the gambia: A case-study in model-based geostatistics. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 51(4):493–506, 2002.

[2] Mengyang Gu, Jesus Palomo, and James O Berger. Robustgasp: Robust gaussian stochastic process emulation. 2016. R package version 0.5.

[3] Peter Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[4] Marc C Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[5] Paulo Ribeiro and Peter Diggle. Analysis of geostatistical data. 2016. R package version 1.7-5.2.

[6] M. Thomson, S. Connor, U. D Alessandro, B. Rowlingson, P. Diggle, M. Cresswell, and B. Greenwood. Predicting malaria infection in gambian children from satellite data and bednetuse surveys: the importance of spatial correlation in the interpretation of results. *American Journal of Tropical Medicine and Hygiene 61*, pages 2–8, 1999.