Analysis Dengue Fever and Malaria Cases using Generalized Multivariate Conditional Autoregressive Model

A Sofro
Mathematics Department, Universitas Negeri Surabaya, East Java, Indonesia
E-mail: ayuninsofro@unesa.ac.id

Abstract. An important topic in epidemiology is dengue fever. This disease is a mosquito-borne infection found in tropical and sub-tropical regions around the world. Another disease infected by mosquito is malaria. The developed methods still offer a lack information how to accommodate a cross correlation between two diseases for different cities. Hence we will provide a natural framework involving a cross correlation for multivariate spatial effect, i.e. generalized multivariate conditional autoregressive model (GMCAR). We also focus on analysing the number of infected cases for two diseases in East Java Indonesia as the response variables. Therefore, in this paper we will propose a multivariate count data modelling with GMCAR to analyse the data. The result shows that the proposed model provides a better performance compare a model without cross correlation.

1. Introduction
Dengue fever and malaria are diseases which infected by mosquitoes and typically occurred in tropical and subtropical regions. The dengue virus causes dengue fever can develop into dengue hemorrhagic fever. Africa, the Americas, the Eastern Mediterranean, and South-east Asia and the Western Pacific regions and more 100 countries are endemic of the disease.

Meanwhile, another a mosquito-borne infectious disease of humans caused by parasitic protozoans is malaria. Similar with dengue fever, coma can develop in malaria cases. Areas of the Americas, many parts of Asia and Africa is also becoming high-risk regions for the disease.

Because both of two diseases come from mosquitoes, hence it is more reasonable if we analyze two diseases simultaneously. Hence, there are multiple diseases in a different area which need to analyze involving a correlation in multiple factors. The correlation is expected to be able to accommodate several types of correlations, i.e. the spatial correlation for each disease across regions, dependence among multiple diseases within the same region and cross spatial correlation among multiple diseases in different regions.

The diseases have been analyzed massively recent year. Many of the researcher have tried to explore either dengue fever or malaria cases using visualization data, involving climate factor, see [1] [2], [3], [4], [5] and [6]. The investigation has improved with combining malaria disease in the analysis, see [7] and [8]. Unfortunately, most of their approaches provide a lack of information in terms of cross-correlation within and between two diseases.
The spatial correlation in the same region has been developed since 1995, see [9], [10], [11], [12], [13], [14], [15]. Meanwhile, the correlation within and between two disease in the different regions (cross correlation) has been provided by [16].

We will focus on analysing the number of infected cases due to two diseases in each city as dependent variables. Thus, in this paper, we will discuss how to model the diseases using multivariate Poisson regression with accommodating the cross correlation. We begin the investigation from the method with the simple model to the complex for multivariate cases.

2. Method
In this section, we will explore the some approaches start from the simple one. i.e Poisson regression.

2.1. Poisson Regression Model
In modelling disease incidence rates, there are a definition which needs to understand. The most common indicator used for comparing regional death rates is the standardized mortality ratio (SMR) in place $i$ which is defined by

$$\text{SMR}_i = \frac{z_i}{E_i},$$

where $z_i$ is observed number of deaths in place $i$ and $E_i$ is the expected number of deaths based on reference mortality rates applied to the regional demographic structure. Also $E_i$ is defined as

$$E_i = \sum_g r_g \times p_{g,i},$$

where $r_g$ is a standard mortality rate (i.e national mortality rate) of demographic group $g$ and $p_{g,i}$ is regional population size specific to demographic group $g$ in place $i$. The demographic group $g$ is usually determined by age or sex-age attributes.

In a generalized linear model (GLM), there is a link function $h$ connecting mean from the dependent variable with independent variables ($U$), i.e.

$$E\left(\frac{z_i}{E_i}\right) = h\left(\frac{\mu_i}{E_i}\right) = U_i^T \beta$$

So we can write the above equation as follows

$$h\left(\frac{\mu_i}{E_i}\right) = \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ... + \beta_k U_{ik}, \quad i = 1, ..., n, \quad (1)$$

where $n$ is the number of observations and $k$ is the number of predictors (covariates). Function $h$ is called the link function. We can also rewrite (1) as

$$\frac{\mu_i}{E_i} = h^{-1}(U_i^T \beta)$$

There are two link functions that can be used in Poisson regression. The first is the identity link and the second is the log link. The form of identity link function as follows :

$$\frac{\mu_i}{E_i} = U_i^T \beta.$$
The log link is
\[
\log \left( \frac{\mu_i}{E_i} \right) = U_i^T \beta.
\]

If we use the log link, the relationship between mean and independent is described below.
\[
\mu_i = E_i e^{U_i^T \beta}.
\]

The log link function is a more popular one because the link guarantees that the value of independent variable is non negative. In term of Poisson regression, the common link function is the log link function and the model is:
\[
\mu_i = E_i \exp(\beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ... + \beta_k U_{ik}) = E_i e^{U_i^T \beta}. \tag{2}
\]

Hence, the probability density function of Poisson regression can be written as follows
\[
P_r(z_i; \beta) = e^{-\mu_i} \frac{\mu_i^{z_i}}{z_i!}.
\]

where \( \mu_i = E_i \exp(U_i^T \beta) \) is the mean and \( \beta \) are unknown parameters. Meanwhile, the mean and the variance for Poisson regression model is defined as
\[
\mu_i = E_i \exp(U_i^T \beta) = Var(z_i).
\]

The likelihood function of Poisson regression is
\[
L(\beta \mid z) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{z_i}}{z_i!}. \tag{3}
\]

2.2. Model of Neighbourhood for Poisson Regression

The Poisson regression with neighbourhood-based spatial effect can be specified as follows
\[
\begin{align*}
z_i \mid s_i & \sim \text{Poisson}(\mu_i) \\
\mu_i & = E_i \exp(U_i^T \beta + s_i) \\
\log(\mu_i) & = \log(E_i) + U_i^T \beta + s_i, \quad i = 1, ..., n
\end{align*}
\]

where the random effect \( s_i \) is assumed to have general conditional autoregressive structure which can be defined as
\[
s_i \mid s_j \sim N(\sum_{i \sim j} b_{ij} s_j, \sigma_i^2).
\]

The component \( b_{ij} \) is the weight of each other observation on the mean of \( s_i \) and \( \sigma_i^2 \) is a variance for \( s_i \). As an example, if state \( i \) has \( m \) neighbours and \( b_{ij} = \frac{1}{m} \) for every state that is a neighbour and 0 otherwise, then the conditional mean of a state’s observation is the mean of all neighbours’ observations. Based on Brook’s lemma, we can write the joint distribution as follows
\[
s \sim N(0, [\tau(D - \alpha W)]^{-1}), \tag{4}
\]

where \( D = \text{Diag}(m_i), m_i \) is the number of neighbors of region \( i \) and \( B = D^{-1}W \). Meanwhile \( W \) denotes the adjacency matrix (\( w_{ii} = 0 \) and \( w'_{ij} = 1 \) if \( i \sim j \) or shared a common boundaries, 0 otherwise). To fulfill requirement of non singular covariance matrix, we define \(|\alpha| < 1\).
2.3. Multivariate Conditional Autoregressive (MCAR) Model for multivariate Poisson Regression

In this section, we will extend the model in equation (4). Therefore, the model can be expressed as follows

\[ z_{ip} \mid s_{ip} \sim \text{Poisson}(\mu_{ip}) \]  
\[ \mu_{ip} = E_i \exp(U_{ip}^T \beta_p + s_{ip}) \]  
\[ \log(\mu_{ip}) = \log(E_{ip}) + U_{ip}^T \beta_p + s_{ip}, \quad i = 1, \ldots, n, \quad p = 1, \ldots, P \]

where \( n \) is the number of area and \( P \) is the number of disease. Meanwhile \( s \) is assumed to have multivariate conditional model (MCAR) which is the extension model from CAR model in the previous subsection. (Carlin and Banerjee, 2003; Gelfand and Vounatsou, 2003.) The specified model of MCAR(\( \alpha, \Lambda \)) can be written as following

\[ s \sim N(0, [(D - \alpha W) \otimes \Lambda]^{-1}). \]  

(6)

We define \( s = (s_1', \ldots, s_p') \) and \( s_k = (s_{1k}, \ldots, s_{nk})' \) with \( k = 1, \ldots, p \) and \( \Lambda \) is positive definite covariance which accommodate nonspatial correlation among variables at any region. In this case, let \( \alpha \in (-1, 1) \), \( D = \text{Diag}(m_i) \) and \( B = D^{-1} W \), \( W \) denotes the adjacency matrix (\( w_{ii} = 0 \) and \( w_{ij} = 1 \) if \( i \sim j \), 0 otherwise).

To understand easily, we give an illustration how the model work. Suppose two diseases in each county, \( p = 2 \) and define \( s_1' = (s_{11}, \ldots, s_{n1}) \) and \( s_2' = (s_{12}, \ldots, s_{n2}) \). MCAR(\( \alpha, \Lambda \)) model has covariance matrix that can be written as

\[ \Sigma = \begin{pmatrix} (D - \alpha W)\Lambda_{11} & (D - \alpha W)\Lambda_{12} \\ (D - \alpha W)\Lambda_{21} & (D - \alpha W)\Lambda_{22} \end{pmatrix}^{-1} \]  

(7)

Carlin and Banerjee (2003). Meanwhile, the covariance matrix of MCAR(\( \alpha_1, \alpha_2, \Lambda \)) model has been specified as

\[ \Sigma^{-1} = \begin{pmatrix} R_1^T R_1 \Lambda_{11} & R_1^T R_2 \Lambda_{12} \\ R_2^T R_1 \Lambda_{21} & R_2^T R_2 \Lambda_{22} \end{pmatrix} = \begin{pmatrix} R_1^T & 0 \\ 0 & R_2^T \end{pmatrix} (\Lambda \otimes I_{n \times n}) \begin{pmatrix} R_1 & 0 \\ 0 & R_2 \end{pmatrix} \]  

(8)

where \( R_j^T R_j = D - \alpha_j W \) and \( |\alpha_j| < 1 \) for \( j = 1, 2 \). We define \( R_j \) as the Cholesky decomposition of \( D - \alpha_j W \) and also \( R_j \) is an upper triangular matrix. In this case, we need to take \( |\alpha_1| < 1 \) and \( |\alpha_2| < 1 \) to produce positive definite covariance matrix. For that reasons, this approach provide some challenges. One of them is difficulty to define the covariance matrix. Hence, in the next subsection we will explore the generalized model using directly specify the covariance matrix.

2.4. Generalized Multivariate Conditional Autoregressive (GMCAR) Model For Multivariate Poisson regression

The multivariate model in this subsection has the same typical with equation (5). The difference is how to define the covariance matrix. To have a good understanding of model, we offer an illustration in two different variables dependent, say \( p = 2 \). Suppose the conditional distribution \( s_1 | s_2 \sim N(As_1, [(D - \alpha_1 W)\tau_1]^{-1}) \) and \( s_2 \sim N(0, [(D - \alpha_2 W)\tau_2]^{-1}) \). We define \( \alpha_1 \) and \( \alpha_2 \) as a smoothing parameters for the two types of dependent variables. Since we have \( E(s_1 | s_2) = As_1 \), where \( A = \eta_0 I + \eta_1 W \) and \( \text{Var}(s_1 | s_2) = [(D - \alpha_1 W)\tau_1]^{-1} \). Then, we assume that its element of \( A \) matrix are

\[ a_{ij} = \begin{cases} \eta_0 & \text{if } j = i \\ \eta_1 & \text{if } j \in n_i, \text{ region } i \text{ and } j \text{ shared a common boundary} \\ 0 & \text{otherwise} \end{cases} \]  

(9)
where $\eta_0$ and $\eta_1$ are the bridging parameters associating $s_{i1}$ with $s_{i2}$ and $s_{j2}$, $j \neq i$ respectively. We denote the model as GMCAR($\alpha_1, \alpha_2, \eta_0, \eta_1, \tau_1, \tau_2$). Hence, if model without involving the cross correlation ($\eta_1 = 0$) and $\alpha_1 = \alpha_2 = \alpha$ then GMCAR($\alpha, \eta_0, \eta_1 = 0, \tau_1, \tau_2$) is exactly the same as MCAR($\alpha, \Lambda$).

### 3. Result and Discussion

The data consist the numbers of infected cases due to dengue fever and malaria each city at East Java Indonesia in 2010. The secondary data is taken from statistics center East Java. East Java province has 29 districts and 9 cities, thus the total of data is 38 observations. The summary of the data is the minimum cases for dengue fever and malaria are 18 and 0 respectively. Meanwhile the maximum cases of dengue fever is 3379 and for malaria is 428 cases.

In terms of selection a proper order two diseases in GMCAR, we need to obtain several steps. The first is calculate estimated of $\hat{s}_{i1} = \log(z_{i1}|E_{i1})$ and $\hat{s}_{i2} = \log(z_{i2}|E_{i2})$. The second is decide the linearity of the conditional GMCAR (say, dengue|malaria),

$$E(s_1|s_2) = A(s_2) = (\eta_0 I + \eta_1 W)s_2.$$  \hspace{1cm} (10)

The third is obtaining $\hat{\eta}_0$ and $\hat{\eta}_1$ by minimizing $(\hat{s}_1 - A(\hat{\eta}_0, \hat{\eta}_1)s_2)'(\hat{s}_1 - A(\hat{\eta}_0, \hat{\eta}_1)s_2).$ Finally, we plot $A(\hat{\eta}_0, \hat{\eta}_1)s_2$ versus $s_1$. Thus, we repeat this process for the reverse order (here, malaria|dengue). Meanwhile, we can explore plot log of SMR and its predicted to assist in selecting a proper order by involving sample correlation of dengue|malaria and malaria|dengue in Figure 1. Hence, in this case, the data provides more evidence to choose linearity dengue|malaria due to providing the highest sample correlation.

In order to understand the performance of several models, we use hierarchical modeling extension of AIC called Deviance Information Criterion (DIC). DIC is based on the posterior distribution of deviance statistic which has a formula as follows

$$DIC = \hat{D} + PD$$

where $\hat{D}$ is denoted as the posterior expected deviance. Meanwhile, $PD$ is defined as the effective number of parameters.

In Table 1, it shows the value of DIC form some different models. The first model is GMCAR($\alpha_1, \alpha_2, \eta_0, \eta_1, \tau_1, \tau_2$). The model provides a full model with conditioning order

![Figure 1. From leftside to rightside plot dengue|malaria with sample correlation 0.4454 and malaria|dengue with sample correlation 0.3643 respectively.](image)
The second model is GMCAR ($\eta_1 = 0$) or we can say as MCAR($\alpha, \Lambda$). The third model is the reverse order from full model where the conditional order is malaria|dengue. From the table, it shows that full model (GMCAR full) provides the smallest of DIC values among all of the comparison models. It means that model with conditional dengue|malaria offers the best model. The reverse order full model is also providing a relatively good performance as well comparing the reduced model (GMCAR ($\eta_1 = 0$)) as we expect.

### Table 1. Model Comparison using DIC Statistics, Dengue Fever and Malaria analysis in East Java Indonesia

| Model                        | $D$   | $PD$  | DIC  |
|------------------------------|-------|-------|------|
| GMCAR (full)                 | 247.70| 5.392 | 253.10|
| GMCAR ($\eta_1 = 0$)         | 247.30| 11.010| 258.30|
| GMCAR (full, reverse order)  | 242.50| 11.730| 254.20|

4. Conclusion
In this paper, involving a spatial effect via GMCAR for modelling multivariate diseases is more appropriate. It is because the method is involving the bridging factor between multivariate responses.

### References

[1] Fareed N, Ghaffar A and Malik TS 2016 Spatio-Temporal Extension and Spatial Analyses of Dengue from Rawalpindi, Islamabad and Swat during 2010 - 2014 Climate vol. 4(2) pp.23

[2] Mustapha E O, Emeribe A U and Nasir I A 2017 Survey of malaria and anti-dengue virus IgG among febrile HIV-infected patients attending a tertiary hospital in Abuja, Nigeria HIV/AIDS - Research and Palliative vol. 9 pp. 145-151

[3] Lin H, Tian L, Yang W, Sun J and Liu Q 2009 Time series analysis of dengue fever and weather in Guangzhou, China BMC Public Health vol. 9 pp. 395

[4] Abbasi A, Abbas K, Arooj S, Habib N, Aziz W and Ashaq A 2016 Dengue Fever: A Statistical Analysis Regarding Awareness about Dengue among University Students in Azad Kashmir Journal of Healthcare Communications ISSN 2472-1654 vol. 2 no. 1: 1

[5] Morin C W, Comrie A C, and Ernst K 2013 Climate and Dengue Transmission: Evidence and Implications Environmental Health Perspect

[6] Mutheneni S R, Morse AP, Caminade C, Upadhyayula S M 2017 Dengue burden in India: recent trends and importance of climatic parameters Emerging Microbes and Infections vol. 6, pp. e70

[7] Das S and Abha T 2016 Predictive analysis of dengue and malaria International Conference on Computing, Communication and Automation (ICCCA) IEEE

[8] Rao MR, Padhy RN and Das MK 2016 Prevalence of dengue viral and malaria parasitic co-infections in an epidemic district, Angul of Odisha, India: An ecoepidemiological and cross-sectional study for the prospective aspects of public health, Journal of Infection and Public Health vol. 9 issue 4 pp. 421-428

[9] Besag J and Kooperberg C 1995 On Conditional and Intrinsic Autoregressions Biometrika vol. 82 pp 733-746

[10] Knorr L and Besag J 1998 Modelling risk from a disease in time and space Statistics in Medicine vol. 17 pp 2045-2060

[11] Sun T, Kim H and He Z Q 2000 Spatio-temporal interaction with disease mapping, Statistics in Medicine vol. 19 pp 2015-2035

[12] McNab Y C and Dean C B 2001 Autoregressive spatial smoothing and temporal spline smoothing for mapping rate Biometrics vol. 57 pp 73-85

[13] Bernardinelli L, Clayton D and Montomoli C 1995 Bayesian estimates of disease maps: how important are priors?Statistics in Medicine vol. 14 pp. 2411-2431

[14] Wall M M 2004 A close look at the spatial structure implied by the CAR and SAR models Journal of Statistical Planning and Inference vol. 121 pp. 311-324

[15] Renato and Krainski E 2009 Neighborhood Dependence in Bayesian Spatial Models Biometrical Journal vol. 51 pp. 851869

[16] Jin X, Carlin B P and Banerjee S 2005 Generalized hierarchical multivariate CAR models for areal data Biometrics vol. 61(4) pp. 950-61