Austrian Journal of Statistics
July 2021, Volume 50, 36–52.
http://www.ajs.or.at/
doi:10.17713/ajs.v50i4.1097

A Generalized Estimating Equations Approach for Modeling Spatially Clustered Data

Nasrin Lipi, Mohammad Samsul Alam and Syed Shahadat Hossain
Institute of Statistical Research and Training, University of Dhaka, Bangladesh

Abstract

Clustering in spatial data is very common phenomena in various fields such as disease mapping, ecology, environmental science and so on. Analysis of spatially clustered data should be different from conventional analysis of spatial data because of the nature of clusters in the data. Because it is expected that the observations of same cluster are more similar than the observations from different clusters. In this study, a method has been proposed for the analysis of spatially clustered areal data based on generalized estimating equations which were originally developed for analyzing longitudinal data. The performance of the model for known clusters is tested in terms of how well it estimates the regression parameters and how well it captures the true spatial process. These results are presented and compared with the conditional auto-regressive model which is the most frequently used spatial model. In the simulation study, the proposed generalized estimating equations approach yields better results than the popular conditional auto-regressive model from the both perspectives of parameter estimation and spatial process capturing. A real life data on the vitamin A supplement coverage among postpartum women in Bangladesh is then analyzed for demonstration of the method. The existing divisional clustering behavior of vitamin A supplement coverage in Bangladesh is identified more accurately by the proposed approach than that by the conditional auto-regressive model.

Keywords: spatial, GEE, areal, CAR.

1. Introduction

In general, most development indicators of any country are measured and reported at national scale. Although limited attention to urban versus rural differences is apparent, local-area variations have not been paid so much attention to. For monitoring and evaluation purpose, planners often require regionally disaggregated indicators. These, too, are expected to be smoothed for the spatial effect because geographic or administrative regions are, in reality, not independent of its nearest neighbors. Although spatial thinking and the use of regionally disaggregated data in the geological and physical sciences have grown rapidly, their implementation in the development research has lagged. Spatial statistical analysis, in recent years, has become very popular due to a rapid development of current knowledge regarding innovations in geospatial data, spatial statistical methods, the integration of data and models and the advances in technology.

Spatial data come as a realization of a stochastic process where the index parameter is space. The statistical methods that are used to analyze spatial data compose the branch spatial statistics. They are different according to the three types of spatial data: geo-statistical or point referenced data, areal or lattice data and point pattern data. For modeling areal data, conditional auto-regressive (CAR) and simultaneous auto-regressive (SAR) are two popular models. In terms of estimation and interpretation of model parameters, CAR is preferred over SAR model, and it also gives minimum mean square prediction error (Cressie 1993). CAR model analyzes the data assuming conditional dependence between observations that is observation of a given site is dependent only on the observations of its neighboring sites.

Spatially clustered data, defined by Knox (1989) as 'a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance, is a very common feature in real life research related to disease mapping, ecology, socio-demographic study and many other disciplines. Often such clusters have high degree of similarity of the observation within cluster and low similarity between clusters. Spatial clustering in the data plays an important role in exploratory data analysis and building statistical models (Jacquez 2008). These kind of analysis is a regular necessity in policy level of any country where regional and sub-regional planning are in place. For example, Bangladesh, has seven administrative divisions and each of the divisions are subdivided into number of districts giving 64 districts in total in the whole country (note that recently the number of divisions is increased to eight, but neither the shape file nor the socio-demographic characteristics are available at the new division levels). Almost all policy level interventions regarding development, education and health are executed at the division levels and they are re-distributed to the districts within the divisions. The development indicators of the districts are, therefore, presumed to be spatially clustered. For spatially clustered data, clusters are independent whereas in CAR model the independence between clusters is not ensured. As a result, the expected prediction of CAR in the border of two clusters will be misleading because there are some observations in the data which are not actually dependent on their neighbors. But CAR model takes their neighborhood into consideration as they are adjacent to each other, therefore, addresses the spatial auto-correlation between them. Lawson and Denison (2002) noted that by nature spatial clusters are not representative of the entire study region. Therefore it is inappropriate to assume a stationary covariance structure for entire study region. As a result, it seems reasonable to think that the CAR model may fail to capture the true spatial process of a spatial clustered data, and may give an inaccurate prediction for modifiable areal unit.

Generalized estimating equations (GEE) is an approach to model correlated data in longitudinal study. It is an extension of generalized linear model (GLM) first introduced by Liang and Zeger (1986). In longitudinal study, the responses are measured repeatedly over time. The GEE allows a correlation structure among the responses. In GEE, the data set is split into some clusters with correlation within clusters, while correlations between clusters are assumed to be zero. In the estimation method, the within cluster correlation is dealt with by assuming a working correlation structure. In GEE literature, there exist different kind of working correlation structures such as exchangeable, auto-regressive of order one, toeplitz, unstructured etc. Inappropriate choice of correlation structure in GEE will lead to inefficient parameter estimation (Hin and Wang 2009). Therefore, to apply GEE on spatial data, we need to choose an appropriate working correlation structure compatible with spatial autocorrelation. Spatial auto-correlation is defined by the 'first law of geography' that 'neighboring regions are likely to be more correlated than the distant regions' i.e. correlation between two regions decreases with the increase of distance. But none of the working correlation structures usually used for GEE (exchangeable, autoregressive etc.) relates or fits to this above kind of spatial auto correlation. That clearly indicates that application of GEE to spatially clustered data for estimation purpose would need working correlation structure appropriate for the spatially clustered data.

There is only a few studies that actually used GEE to analyze spatial data. It was first developed by Albert and McShane (1995). The author proposed the approach for modeling spatial

location and subject specific covariates. The GEE was applied concentrating on the marginal mean structure, treating the spatial correlation as nuisance. The goal was to make inference about marginal mean in the presence of spatial correlation. They used semivariogram model for characterizing the correlation structure spatial data sets. Following the first application several authors applied GEE in spatial data (Gotway and Stroup 1997; Gumpertz, Pye *et al.* 2000; Augustin, Kublin, Metzler, Meierjohann, and von Wühlisch 2005). In these studies, authors applied GEE on geostatistical data and for small sample size considering the whole study region as one cluster. Clustering of the data is neither naturally present nor artificially imposed in these studies. For large data, computation will be very difficult assuming the whole study region as one cluster. Carl and Kühn (2007) first tested the performance of GEE method for data of regular gridded maps and large sample size. F Dormann, M McPherson, B Araújo, Bivand, Bolliger, Carl, G Davies, Hirzel, Jetz, Daniel Kissling *et al.* (2007) conducted a review study of the methods of accounting spatial auto-correlation where GEE was used in the same way as Carl and Kühn (2007) on regular lattice. They used GEE method by considering the entire area as one cluster and making up artificial clusters combining only a few grid cells. The clusters are made arbitrarily by their developed function. The function made clusters of different sizes such as: 2×2, 3×3 and 4×4 clusters. Maximum cluster size can not exceed 4×4. Therefore, the potentially existing correlations among sites which belong to different clusters are neglected. Furthermore, cluster size can be more than 4× 4 grid cells, which is not supported by the function.

The main goal of this study is to apply GEE as a method of modeling areal data. In our study, we will apply GEE on spatially clustered data for estimation of regression parameters. In addition, a working correlation structure appropriate for the spatially clustered data will be suggested and hence a revised process that is assumed to be a better representation of the true process than the observed process is shown using the GEE approach to describe how well our model is capturing the true process. Simulated spatial map describing these revised estimates will be compared with that describing the true process and also with the most popular spatial smoothing method, the conditional auto-regressive (CAR) model. The methodology will be demonstrated for an example data of district wise vitamin A supplement coverage among postpartum women in Bangladesh.

2. The proposed approach

The proposed approach for applying GEE on spatially clustered data will utilize the theories of conditional auto-regressive model (CAR) (Besag 1974) and the generalized estimating equations (GEE) (Liang and Zeger 1986) which are briefly given in the following two sections.

2.1. Conditional auto-regressive model

Consider n non-overlapping sub-areas $S_1, S_2, \ldots, S_n$ of study region $\mathcal{S} \subset \mathbb{R}^2$ that are spatially proximate and which are linked to a corresponding set of response $Y_1, Y_2, \ldots, Y_n$. Then $\mathcal{S} = \{S_1, \ldots, S_n\}$ defines a spatial lattice. Let y_i denote the realization of the response in the i-th sub-area, $i = 1, 2, \cdots, n$. Under the spatial set up, the distribution of the response Y_i of the i-th sub-area is thought to be dependent on the realizations of the j-th ($j \neq i$) sub-area, hence the conditional distribution of the response Y_i given y_j, $j \neq i$ is of interest. Statistical analysis with such kind of data is done under a set of assumptions. The CAR model assumes that this conditional distribution can capture the spatial dependence by defining it through neighborhood structure of the sub-areas. We denote by $\mathbf{W} = \{w_{ij}\}$ the matrix that defines the neighborhood relationship between sub-area S_i and S_j, which can be defined by any of the methods such as common boundary, distance based method etc. A traditional method is to consider w_{ij} as an indicator variable taking value 1 if S_i and S_j share a common boundary and 0 otherwise. For continuous type of data, a wide-spread assumption is the Gaussian or

the auto normal assumption which states that

$$Y_i|y_j, j \neq i \sim N\left(\sum_j b_{ij}y_j, \tau_i^2\right), \tag{1}$$

Further it can be extended for spatial covariate such as for each sites the variable of interest Y_i and a set of $p < n$ explanatory variables $\mathbf{x}_i = (x_{i1}, ..., x_{ip})'$ are observed. The effect of these covariates on the Y_i i.e. large scale effect of the spatial model is a vector $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)'$.

$$Y_i|y_j, j \neq i \sim N\left(\mathbf{x}_i'\boldsymbol{\beta} + \sum_j b_{ij}(y_j - \mathbf{x}_j'\boldsymbol{\beta}), \tau_i^2\right), \tag{2}$$

with parameters b_{ij} and τ_i^2. If the joint distribution of $Y_1, Y_2, \ldots, Y_n$ can be determined from equation (1), we call it a Markov Random Field. This can be obtained through Brook's Lemma as

$$p(y_1, y_2, \ldots, y_n) \propto exp\left\{-\frac{1}{2}\mathbf{y}'\mathbf{D^{-1}}(\mathbf{I} - \mathbf{B})\mathbf{y}\right\}, \tag{3}$$

where $\mathbf{B} = \{b_{ij}\ ;\ i, j = 1, 2, \ldots, n\}$ and $\mathbf{D}$ is diagonal matrix with $D_{ii} = \tau_i^2$. If $\mathbf{D^{-1}}(\mathbf{I} - \mathbf{B})$ is ensured to be symmetric, equation (3) suggests a multivariate normal distribution for $\mathbf{Y}$ with mean $\mathbf{By}$ and variance matrix $\boldsymbol{\Sigma}_y = (\mathbf{I} - \mathbf{B})^{-1}D$. While $\mathbf{B}$ is not symmetric, we set $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$, where $w_{i+} = \sum_{j=1} w_{ij}$ to make $\boldsymbol{\Sigma}_y$ symmetric. Thus, equations (1) and (3) yield

$$p(y_1, y_2, \ldots, y_n) \sim N(\sum_j w_{ij}y_j/w_{i+}, \tau_i^2/w_{i+}), \tag{4}$$

and

$$p(y_1, y_2, \ldots, y_n) \propto exp\{-\frac{1}{2\tau^2}\mathbf{y}'(\mathbf{D_w} - \mathbf{W})\mathbf{y}\}, \tag{5}$$

where $\mathbf{D}_w$ is diagonal with $(D_w)_{ii} = w_{i+}$. From the definition of $\mathbf{D}_w$ and $\mathbf{W}$, it is noticed that $(\mathbf{D}_w - \mathbf{W})\mathbf{1} = \mathbf{0}$, which implies $\boldsymbol{\Sigma}_y^{-1}$ is singular and the distribution in (5) is improper. This improperness in (5) can be remedied by redefining $\boldsymbol{\Sigma}_y^{-1} = (\mathbf{D}_w - \rho\mathbf{W})$ and choosing ρ to make $\boldsymbol{\Sigma}_y^{-1}$ non-singular. Under $\boldsymbol{\Sigma}_y^{-1} = (\mathbf{D}_w - \rho\mathbf{W})$, the full conditional $p(y_i|y_j; j \neq i)$ becomes $N(\rho\sum_j w_{ij}y_j/w_{i+}, \tau_i^2/w_{i+})$. The breadth of spatial association is limited for a proper CAR model (Banerjee, Carlin, and Gelfand 2014).For the model defined in (2), CAR model can be written as,

$$\mathbf{Y} = \mathbf{BY} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\ . \tag{6}$$

If the joint distribution is proper then $\mathbf{Y} \sim N\left(\mathbf{By} + \mathbf{X}\boldsymbol{\beta}, (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}\right)$ which induces the distribution of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{D}(\mathbf{I} - \mathbf{B})')$.

2.2. Generalized estimating equations

Let, in a set up of longitudinal data analysis, y_{ij} and $\mathbf{x_{ij}}$ denote the response and the p × 1 vector of covariates, respectively, at the j-th time for the i-th subject (i = 1, 2,. . . , N, and j = 1, 2,. . . n_i). The marginal model for the response y_{ij} requires specifying marginal mean $\mu_{ij} = E(Y_{ij}|\mathbf{x_{ij}})$ and variance by a generalized linear model as

$$g(\mu_{ij}) = \mathbf{x_{ij}'}\boldsymbol{\beta} \tag{7}$$

and

$$\text{var}(Y_{ij}) = \phi v(\mu_{ij})\ , \tag{8}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is the vector of regression coefficients corresponding to the vector of covariates $x_i = \{x_{i1}, x_{i2}, \ldots, x_{ip}\}$, $g(\cdot)$ is the link function, v is a known variance function, and ϕ is the scale parameter. Thus the covariance matrix of the response is specified as

$$\mathbf{V_i} = \text{cov}(\mathbf{Y_i}) = \phi \mathbf{A_i}^{1/2} \mathbf{R} \mathbf{A_i}^{1/2} \ , \tag{9}$$

where, $\mathbf{A_i}$ is a $n_i \times n_i$ diagonal matrix with elements $\text{var}(Y_{ij}) = \phi v(\mu_{ij})$ as the j-th diagonal element, $\mathbf{R_i}(\boldsymbol{\alpha})$ is the correlation matrix among the outcomes measured at different times for the i-th subject and $\boldsymbol{\alpha}$ is a q-dimensional vector of unknown parameters that completely specifies within subject correlation. By the generalized estimating equations (GEE) approach, estimate of the vector of regression coefficients $\boldsymbol{\beta}$ is obtained by solving the following equations

$$\sum_{i=1}^{N} \mathbf{D}_i' \mathbf{V}_i^{-1} (\mathbf{y_i} - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{10}$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ is the marginal mean vector for the subject i and $\mathbf{D}_i = \frac{\delta \mu_i}{\delta \beta}$. According to Liang and Zeger (1986), the estimation of $\boldsymbol{\beta}$ requires specifying the structure of the working correlation matrix $\mathbf{R_i}(\boldsymbol{\alpha})$. Here, the variance-covariance matrix in equation (10) has block diagonal form, since responses of different subjects are assumed to be uncorrelated. If the parameters in $\mathbf{R}$ are not known, it will be necessary to estimate all of them in an iterative procedure until convergence is achieved. In addition the standard errors of the estimates are computed using the usual sandwitch estimator.

2.3. GEE in case of areal data

Consider a spatial lattice $\mathcal{S} = \{S_{ij} \ ; i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, m_i\}$ in presence of spatial clustering, where $\{S_{ij} \ ; \ j = 1, 2, \ldots, m_i\}$ constitutes the i-th cluster and S_{ij} is the j-th sub-area in the i-th cluster. We assume that sub-areas have high degree of similarity of the observation within cluster and low similarity between clusters. In such setup, we can utilize the GEE approach described in Section 2.2 by considering each of the clusters as individual subject and the sub-area within each cluster as correlated responses of the cluster. The estimation of regression parameter in GEE requires to define the mean model and the structure of working correlation matrix. This paper would consider the situation of continuous data only, hence the form of link function is identity. In spatial context, we define correlation structure for i-th cluster as

$$\mathbf{R}_i = \left\{ \exp(-\hat{\lambda}_i \times d_{i(jk)}) \ ; \ j, k = 1, 2, \ldots, m_i \right\}, \tag{11}$$

where $d_{i(jk)}$ is the Euclidean distance between centers of sites j and k within i-th cluster and $\hat{\lambda}_i$ is a scale function controlling the decay of the dependence of spatial auto-correlation on the distance between units in i-th cluster. Use of the scale parameter λ_i in the above correlation structure is intended to encompass different level of spatial heterogeneity among the units of the i-th cluster. This study assumes that the random variables associated with areal units in the same cluster are positively correlated and the first order Moran's I for the i-th cluster is used as an estimate of λ_i. There are two components in the error of a spatial data. One is spatial effect and another is the random noise. Here variance of $\mathbf{Y_i}$ is decomposed as

$$\text{cov}(\mathbf{Y_i}) = \text{cov}(\boldsymbol{\nu_i}) + \text{var}(\boldsymbol{\epsilon_i}),$$

where $\text{cov}(\boldsymbol{\nu_i})$ is the variance of spatial effect which contains covariance term because $\boldsymbol{\nu_i}$ is spatially correlated and $\text{var}(\epsilon_i)$ is a diagonal matrix containing variance of random noise in the i-th cluster. $\text{cov}(\boldsymbol{\nu_i})$ can be defined as

$$\text{cov}(\boldsymbol{\nu_i}) = \mathbf{A_i^{\frac{1}{2}}} \mathbf{R_i} \mathbf{A_i^{\frac{1}{2}}} \ ,$$

where $\mathbf{A_i}$ is a $n_i \times n_i$ diagonal matrix which contains $\text{var}(\nu_{ij})$ as the jth diagonal element and $\mathbf{R_i}$ is the correlation matrix among the units of the i-th cluster which is defined in equation (11). Now, $\text{cov}(\mathbf{Y_i})$ is defined as

$$\begin{aligned} \mathbf{V_i} = \text{cov}(\mathbf{Y_i}) &= \text{cov}(\boldsymbol{\nu_i}) + \text{var}(\boldsymbol{\epsilon_i}) \\ &= \mathbf{A_i^{\frac{1}{2}}} \mathbf{R_i} \mathbf{A_i^{\frac{1}{2}}} + \sigma_e^2 \mathbf{I_{n_i}} \ . \end{aligned} \tag{12}$$

As the estimating equations in equation (10) require an estimate of $\mathbf{V_i}$, this approach utilized the idea of analysis of variance. Here, $\mathbf{A_i}$ is estimated by the residual variance of the analysis of variance of $\mathbf{Y}$ with cluster as factor because units within a cluster are correlated but between clusters the units are independent and σ_e^2 is estimated by the between cluster variance of $\mathbf{Y_i}$. Finally, regression parameters are estimated by solving equations (10) with estimated $\mathbf{V_i}$.

2.4. Mapping of the fitted value

Estimated residuals $\mathbf{e_i} = \mathbf{y_i} - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ from GEE model are correlated because it contains both the spatial effect and the random error term. To extract spatial effect from $\mathbf{e_i}$ we will use the residual analysis techniques of longitudinal analysis. We will de-correlate the residuals to get the error term. This can be done by Cholesky decomposition. Given an estimate of the covariance matrix for the residuals, $\hat{\mathbf{V}}_\mathbf{i}$, the Cholesky decomposition is defined as

$$\hat{\mathbf{V}}_\mathbf{i} = \mathbf{L_i}\mathbf{L_i'},$$

where $\mathbf{L_i}$ is a lower triangular matrix. The residuals are made uncorrelated by

$$\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}} = \mathbf{L_i^{-1}} \mathbf{e_i} = \mathbf{L_i^{-1}} \left(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}} \right) \ . \tag{13}$$

The estimate of the random noise is given by $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}$ in equation (13). Spatial effect is estimated by $\hat{\boldsymbol{\nu}}_{\boldsymbol{i}} = \mathbf{e_i} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{i}}$. Finally, the fitted value for the model is calculated by $\hat{\mathbf{y}}_\mathbf{i} = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\nu}}_{\boldsymbol{i}}$.

3. Simulation study

To assess the performance of our proposed model of the GEE approach for estimating the regression parameter $\boldsymbol{\beta}$ in equation (7) in case of spatial clustering in data, we conducted a Monte Carlo simulation. Since, in presence of spatial effect, it is almost an obvious choice to use CAR model, we compared the results of GEE approach with those obtained using simple CAR model for the same data.

We have generated artificial data and in order to ensure the generated data to be spatially correlated, we used the procedure described by Dormann (2007). It was also taken under consideration of the simulation to make the generated data having a spatial clustering. The analysis is conducted in R (R Core Team 2017).

The simulation was conducted under a hypothetical spatial scenario of administrative segmentation of Bangladesh. Bangladesh is divided into seven administrative divisions, each of them are also sub-divided into a number of upazillas (sub-areas). All together, there are 458 upzillas in Bangladesh and they are distributed in the 7 divisions as in Table 1.

As a result, our generated dataset represents a hypothetical data collected from 458 upzillas of Bangladesh. On this lattice, the artificial data y_{ij} $(i = 1, 2, \ldots, 7; j = 1, 2, \ldots, m_i)$, were generated as a function of a single predictor x_{ij} which was generated from standard uniform distribution.

In the functional relationship between y_{ij} and x_{ij} as described in Section 2.2, a spatial autocorrelation was introduced as spatially correlated errors.

Table 1: Number of Upazillas in each of the 7 divisions of Bangladesh

Division	i	Number of Upazillas (m_i)
Barisal	1	37
Chittagong	2	85
Dhaka	3	119
Khulna	4	58
Rajshahi	5	66
Rangpur	6	58
Sylhet	7	35

A weight matrix $\mathbf{W}$ was used to convert the random noise to spatially correlated errors $\boldsymbol{\nu_i}$ using weights according to the distance between data points. Let $\mathbf{D} = \{d_{ij}\}$ be the (Euclidean) distance matrix for the distances between upzillas i and j. Then $\mathbf{W} = \{w_{ij}\}$ is calculated by

$$w_{ij} = \begin{cases} \exp(-\lambda \times d_{ij}) \ , & \text{if } i\text{-th and } j\text{-th upazillas are in the same cluster} \\ 0, & \text{Otherwise} \ , \end{cases} \tag{14}$$

where $\lambda(\lambda \geq 0)$ determines the decline of inter-cell correlations in error with increasing inter-upzilla distance. Clearly, the strength of spatial auto-correlation at a given distance increases as the value of λ decreases. Since, the generated data are needed to be spatially clustered in 7 clusters (divisions), the weight matrix $\mathbf{W}$ is constructed as a block diagonal matrix with seven blocks. This $\mathbf{W}$ matrix makes the random noise spatially correlated within each division and independent between the divisions. The vector of independent errors, drawn from the standard normal distribution, was multiplied by Φ, the Cholesky decomposition of $\mathbf{W}$ ($\mathbf{W} = \Phi'\Phi$). By this procedure correlated errors $\boldsymbol{\nu_i} = \{\nu_{ij} \ ; \ j = 1, 2, \ldots, m_i\}$ are generated. This correlated error introduce spatial autocorrelation in our generated data. Further, to introduce an unstructured variation in data an unstructured noise $\boldsymbol{\epsilon_i} = \{\epsilon_{ij} \ ; \ j = 1, 2, \ldots, m_i\}$ drawn from a normal distribution with mean 0 and standard deviation 10 is also used in generating $\mathbf{y_i} = \{y_{ij} \ ; \ j = 1, 2, \ldots, m_i\}$. Finally, y_{ij} is simulated as

$$y_{ij} = 80 - 0.015 \times x_{ij} + 10 \times \nu_{ij} + \epsilon_{ij} \ ; \ i = 1, 2, \ldots, 7; j = 1, 2, \ldots, m_i \ .$$

4. Results of simulation study

To account for different level of spatial correlation, different values 0.05 to 0.55 in steps 0.05, of λ were considered for the simulation. For each scenario, 1000 simulated data were generated under the setup given in section 3. Estimates of the model parameters were obtained using our proposed method and CAR model for each of the randomly generated data sets.

Table 2 presents the efficiency of parameter estimates for 1000 simulations, each with 458 upazillas. Results are shown for averaged estimates of regression parameter, standard error, bias, root mean square error and coverage probability for the two methods, our proposed one and the CAR model .

From Table 2, it is observed that on average GEE approach yields better estimates than its counterpart irrespective of the values of λ. Looking at the values of SE(β_1), it is seen that the efficiency of the estimated parameter is lower for CAR model relative to that observed for the proposed method. The same conclusion is suggested by the values of root mean square error of the estimated parameter. Investigating the values of the coverage probability of 95% confidence interval, it is found that the coverage probabilities are not much different for the two methods when the extent of within cluster spatial association is high. In addition, with

Table 2: Mean, standard errors, bias, root mean square error and coverage probability of 95% confidence interval obtained in case slope parameter β_1 from 1000 simulated datasets

	Method	SE($\hat{\beta}_1$)	Bias($\hat{\beta}_1$)	RMSE($\hat{\beta}_1$)	Coverage(%)
$\lambda = 0.05$	GEE	0.00194	0.00016	0.00194	97.2
	CAR	0.00204	0.00039	0.00208	95.3
$\lambda = 0.1$	GEE	0.00194	0.00018	0.00195	95.9
	CAR	0.002044	0.000420	0.002086	95.3
$\lambda = 0.15$	GEE	0.00195	0.00018	0.00196	95.3
	CAR	0.002046	0.000432	0.002091	95.3
$\lambda = 0.2$	GEE	0.00195	0.00017	0.00196	94.9
	CAR	0.002048	0.000433	0.002093	95.5
$\lambda = 0.25$	GEE	0.00196	0.00015	0.00196	94.3
	CAR	0.002049	0.000426	0.002092	95.7
$\lambda = 0.3$	GEE	0.00196	0.00014	0.00197	93.6
	CAR	0.002049	0.000412	0.002090	95.9
$\lambda = 0.35$	GEE	0.00196	0.00012	0.00197	92.9
	CAR	0.002050	0.000394	0.002087	96.1
$\lambda = 0.4$	GEE	0.00197	0.00010	0.00197	92.5
	CAR	0.002050	0.000371	0.002084	96.3
$\lambda = 0.45$	GEE	0.00197	0.00008	0.00197	91.3
	CAR	0.002050	0.000346	0.002079	96.5
$\lambda = 0.5$	GEE	0.00197	0.00006	0.00198	91.4
	CAR	0.002050	0.000318	0.002075	96.5
$\lambda = 0.55$	GEE	0.00198	0.00003	0.00198	90.9
	CAR	0.00205	0.00028	0.00207	96.7

True parameter value β_1 = -0.015

decrease in within cluster spatial auto-correlation, lesser coverage is observed in case of the proposed GEE approach while the coverage for CAR method is seen to be higher than the nominal level of 95%. However, Results obtained also indicate that, if the scale of dependence of spatial association on the distance between two areal units is small, then the proposed model outperforms the CAR. It is also observed that when the scale of dependence of spatial association on the distance increases the proposed method can give deviated results. This finding reveals that the proposed GEE approach would perform better in case of high level of spatial clustering in data. This is apparently due to the reason that, in case of strong within cluster spatial auto-correlation, the clustering pattern will be captured better by the equation (11), while CAR would consider both the within cluster neighbouring units and the bordering between cluster neighbouring units as equally auto-correlated.

To check whether the proposed approach adequately fits the true data and also to check how well the spatial clustering is captured by the proposed method, we produced choropleth maps of the fitted processes obtained through proposed GEE approach and CAR model for different values of λ. For a comparative understanding on the goodness of fit of each of the approaches, scatter plots of true process versus fitted process are also produced for each values of λ considered in the study. The choropleth maps are shown in Figures 1 through Figure 3. In each of the figures, maps for the true process, fitted process by the GEE approach and fitted process by the CAR model are shown along with the scatter plot of true versus fitted process for each of the methods.

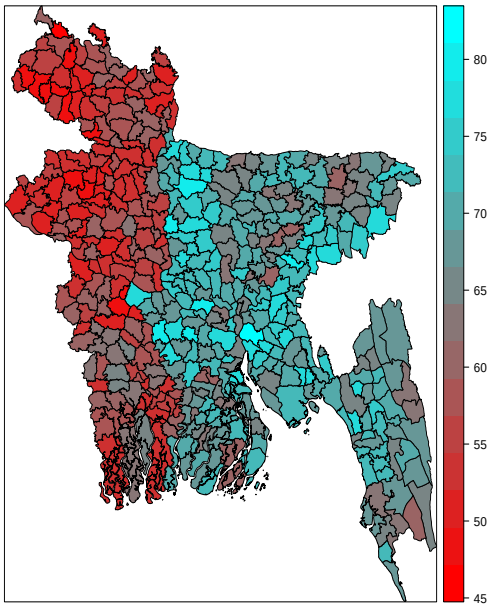
From Figure 1, we can observe that the map for the true process evidently shows spatial clustering, but the map of fitted values obtained from CAR model does not show such clustering. The CAR model being a neighborhood dependent smoothing technique, smooths out the process over the whole area, thus, fails to retain clustering pattern of the true process. Unequivocally, it is noticeable that the fitted process by the proposed GEE approach incorporating the scaled-distance based correlation structure does not ignore the clustered pattern of the original data, and therefore, captured the true process better than CAR does. Similar observations can be made from the maps produced in Figures 1 through Figure 3 for other λ values. This indicates that proposed GEE approach performs better in spatial prediction than CAR model while data are spatially clustered no matter what the scale of dependence of spatial auto-correlation on distance between units is.

Scatter plots of true spatial effect and estimated spatial effect for both model for different λ are shown in graphs (d) of Figures 1 to Figure 3. It is expected that if the used method predicts spatial effect perfectly then the points in the scatter plots would lie on a straight line. That is why deviation from the straight line is used as a measure of the performance of the methods in prediction. The more it deviates from straight line the less the estimates of spatial effect is accurate for the particular model. At all the considered values of λ, it is observed that the GEE approach estimates spatial effect more accurately than the CAR model.

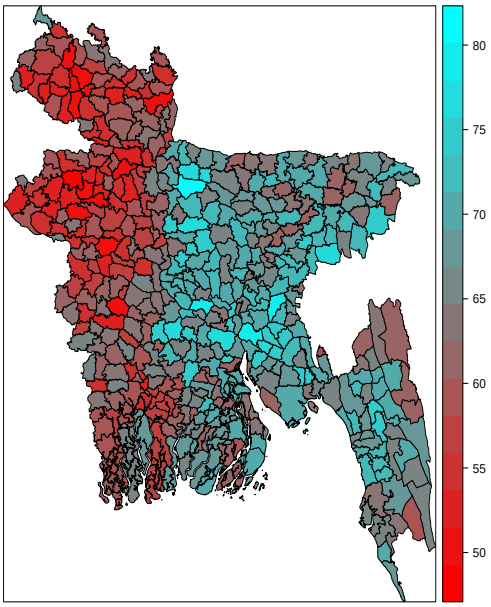
Figure 4 shows bias and root mean square error vs different λ values for CAR and our proposed model. From this plot we can see that for both GEE and CAR bias is decreasing with the increase of λ. Again, for all λ values bias of our proposed GEE method is always lower than CAR. Same behavior is revealed in the root mean square error plot.

5. Application to a real life example

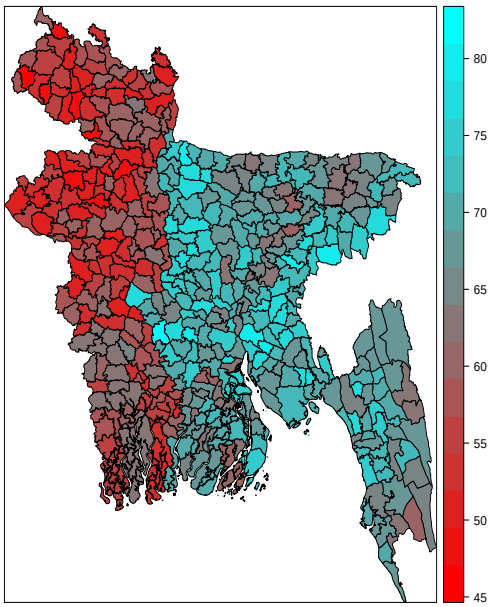
As an example, district wise vitamin A supplement coverage among postpartum women in Bangladesh is used in this study for comparing the proposed GEE approach and existing CAR model. The secondary data collected from "Bangladesh EPI Coverage Survey 2009" (EPI Coverage Evaluation Survey 2009) are used as an example. Note that instead of the smaller administrative units upazilas, districts are considered as sub-areas since the data on vitamin A supplement coverage are not available at the upazila level. As part of a national



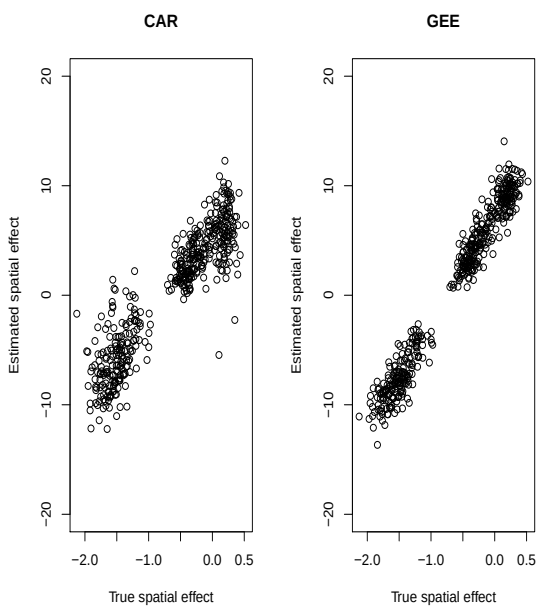
(a) True process



(b) Fitted process by CAR

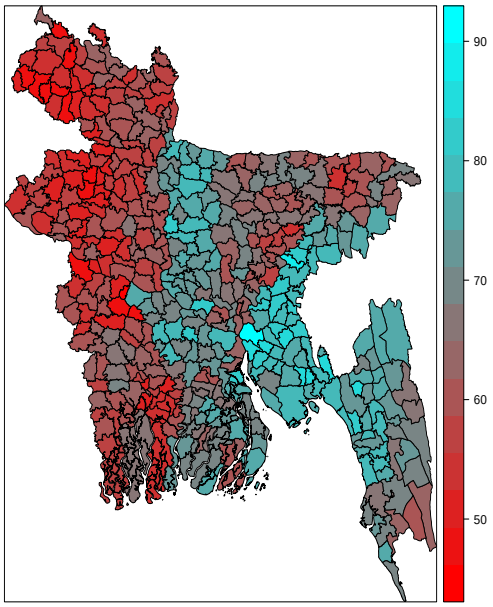


(c) Fitted process by GEE

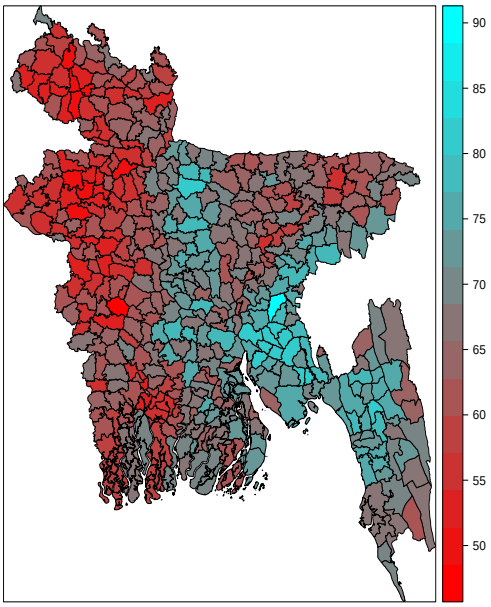


(d) True vs. estimated spatial effect

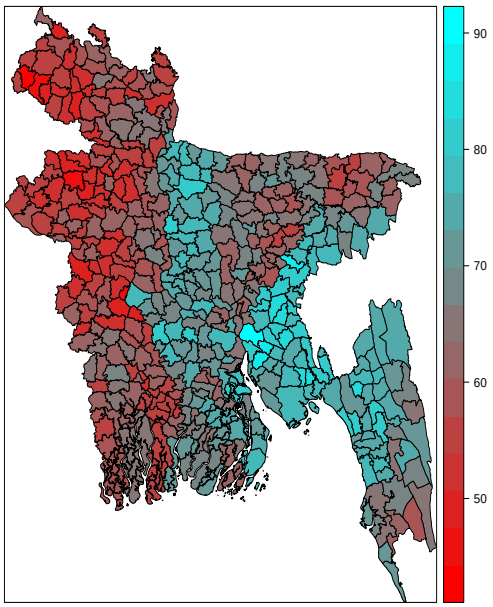
Figure 1: Comparison of predicted maps using CAR and GEE models with the observed process and the scatter plot of true versus fitted process for $\lambda = 0.05$



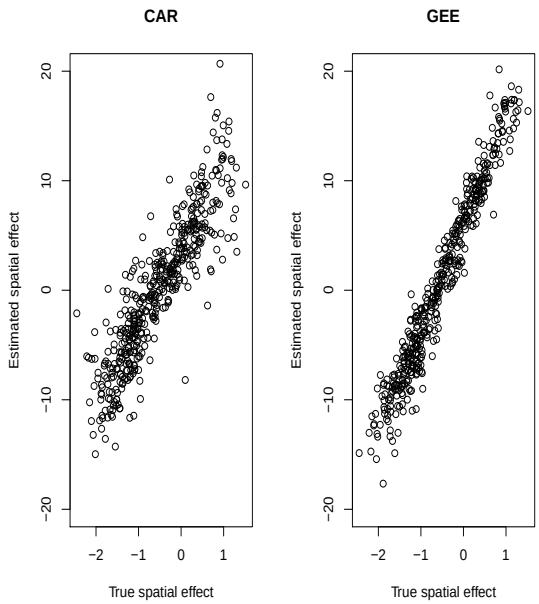
(a) True process



(b) Fitted process by CAR

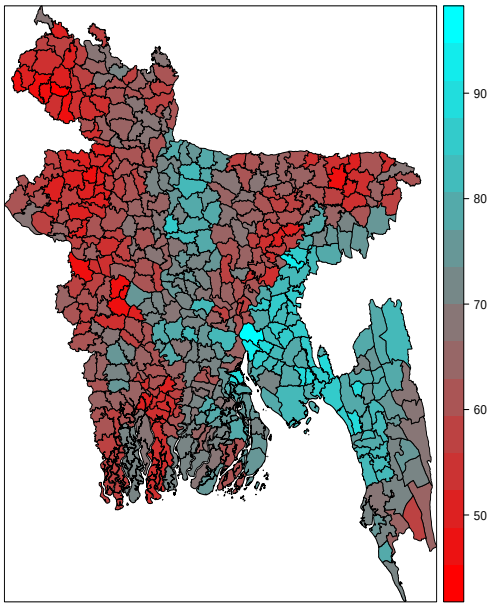


(c) Fitted process by GEE

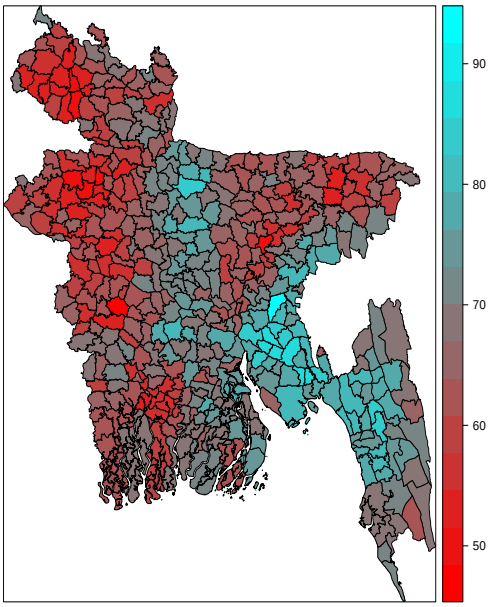


(d) True vs. estimated spatial effect

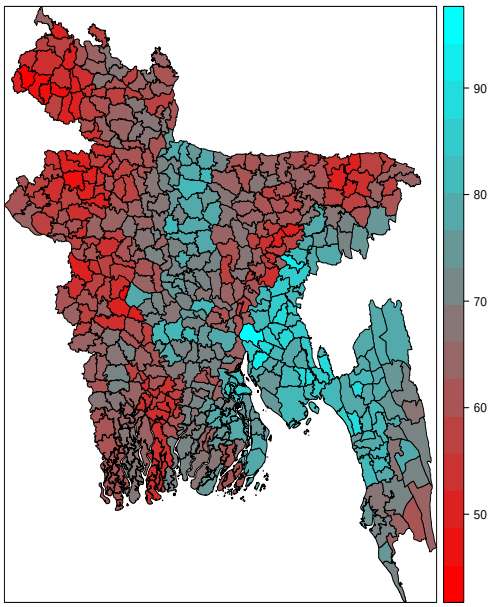
Figure 2: Comparison of predicted maps using CAR and GEE models with the observed process and the scatter plot of true versus fitted process for $\lambda = 0.2$



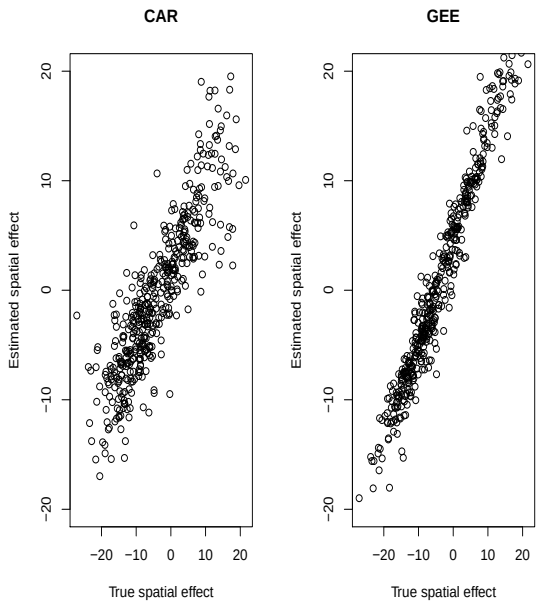
(a) True process



(b) Fitted process by CAR



(c) Fitted process by GEE



(d) True vs. estimated spatial effect

Figure 3: Comparison of predicted maps using CAR and GEE models with the observed process and the scatter plot of true versus fitted process for $\lambda = 0.5$

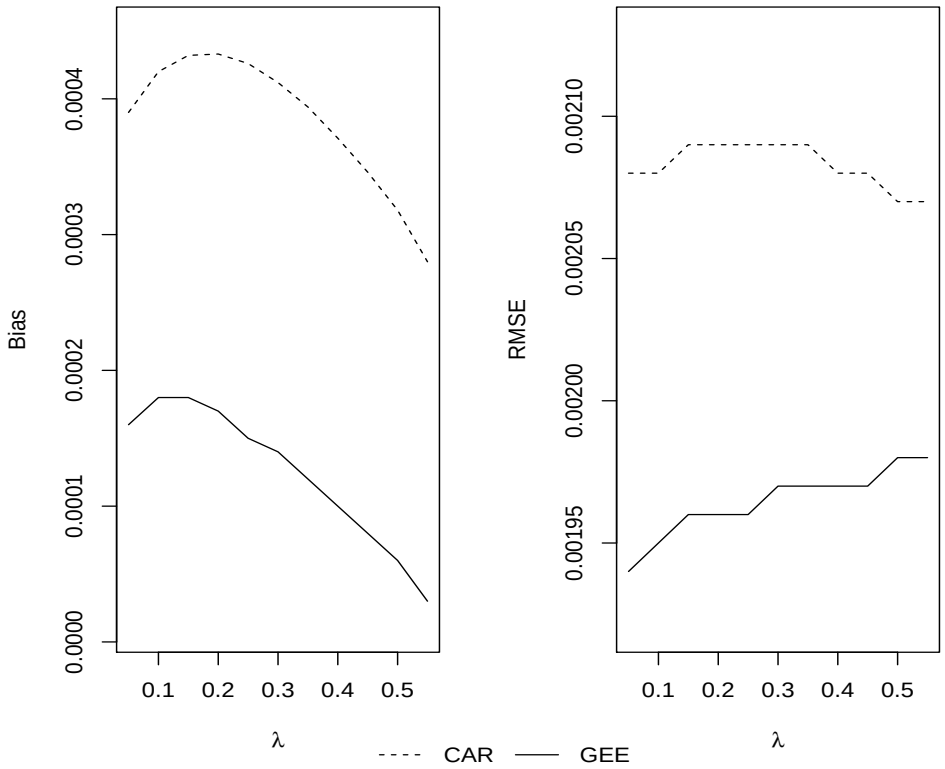


Figure 4: Comparison of Bias and Root Mean Square Error(RMSE) for different λ values

campaign, all postpartum women in Bangladesh are targeted to receive a vitamin A supplement. While due to many reasons this intervention does not reach all the eligible women. However, the government of Bangladesh monitors the progress of this kind of intervention through a coverage survey at a regular time interval. Although the program is a nationwide approach, the progress of the program is measured by the outcome variable, vitamin A supplement coverage rates (VASCR), which differs due to execution of the program conducted at the division level. While the intensity and efficiency of the execution of the program by different division are in practice different, we considered that the VASCR of districts of different divisions to be spatially independent. But with the obvious conviction that VASCR of districts within the same division are spatially correlated, the district estimates of vitamin A supplement coverage rates (VASCR) reported in the publication EPI Coverage Evaluation Survey (2009) are used here as an example that would fit to the described GEE approach. Figure 5 (a) shows the map of observed VASCR data for different districts. Divisions are identified in the map and the division boundaries are marked by dark lines.

Note that the program, being conducted with division wise administrative setup, we consider the VACR of districts of different divisions to be spatially independent and that of districts within the same division are spatially correlated. Under this assumption, the VASCR of district Habiganj in Sylhet division is expected to be more influenced by the VASCR in other three Sylhet division districts than by that in its neighbor districts Brahmanbaria (part of Chittagong division) and Kishoganj (part of Dhaka division). Assuming the VASCR of the three districts Habiganj, Brahmanbaria and Kishoganj to be correlated may not be reasonable as the three division have three different level of priority, intensity and efficiency for the vitamin A supplement program. That is why, we utilized the GEE approach of estimation and prediction for VASCR of each of the districts. For an understanding of how it could have been misleading by assuming that 'all neighbors are likely to be correlated' - we generated

the estimates and prediction using CAR model as well.

Table 3 shows the estimates of the regression parameters (intercept) and their corresponding standard error for both the proposed method and CAR model. Figures 5 (b) and 5 (c) show the comparison of the fitted values by proposed GEE approach and CAR model with the observed data (5 a).

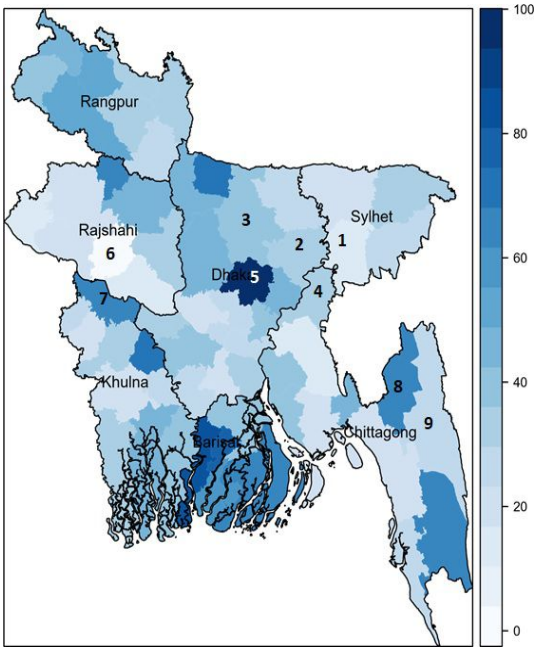
Table 3: Estimates of the regression parameter and respective standard errors in case of vitamin A supplement coverage among postpartum women

Methods	Estimates	Standard Error
GEE	37.375	0.413
CAR	38.236	3.054

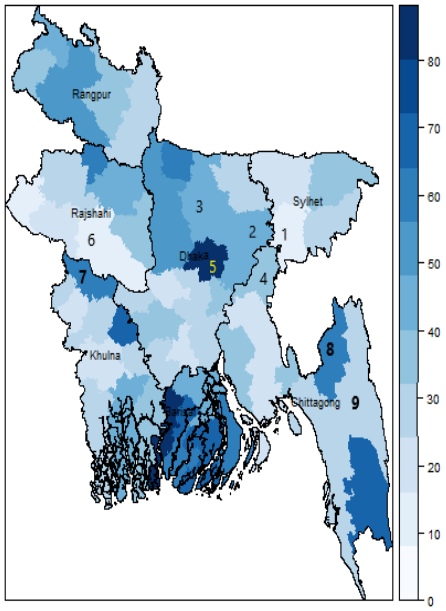
Table 3 reveals that for the vitamin A supplement coverage rate, GEE approach estimates the regression parameter with a standard deviation lower than that for the estimates using CAR model. More importantly, from Figure 5, it is clearly depicted that the GEE approach can capture the clustering more accurately than the CAR model does. It can also be observed from the Figure 5 that the VASCR of district Habiganj (1) in Sylhet has been substantially over estimated by CAR method, while the VASCR in the other Sylhet division districts are moderately low. This has happen evidently due to the reason that the estimation has taken into consideration the spatial correlation between the VASCR of Habiganj (1) with its neighbor districts Kishoganj (2) (part of Dhaka division) and Brahmanbaria (4) (part of Chittagong division). Yet again, the estimated VASCR in Gazipur (5) district has smoothed down to a quite low level from a high observed value. Noteworthy that VASCR is a program driven outcome and expected to be dependent on the strength and tenacity of the program which is executed at the division level. That is why, considering the division as cluster, the GEE approach has provided more realistic estimated VASCR. In each of the two above discussed cases of Habiganj and Gazippur, the GEE approach has given smoothed VASCR which are not contradicting too much with the true picture. Such discrepancies between the two approaches are observed for some other districts bordering the divisions, such as: Natore (6), Kushtia (7), Mymensing (3), Khagrachari (8) and Bandarban (9) districts. This is because CAR model does not take clustering pattern in consideration while predicting the data. On the other hand, GEE approach assumes the clustering and considers spatial clustering and independence between them. For this reason GEE model performs better than CAR for spatially clustered data.

6. Discussion & conclusion

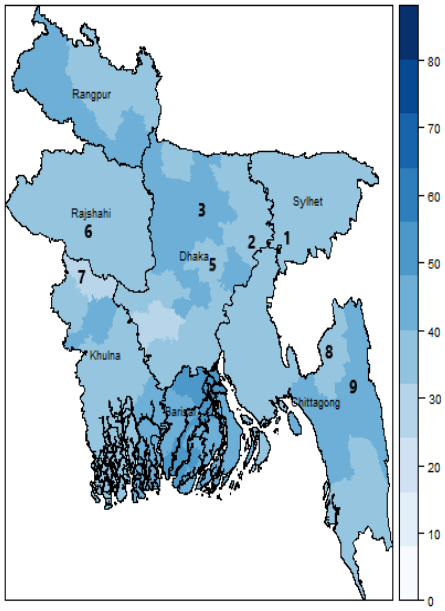
This study has simulated spatially clustered data with known clusters. For known cluster the proposed method performs better than traditional conditional auto-regressive (CAR) model in terms of regression parameter estimates and prediction performance. In spatially clustered data, sites within a cluster are dependent but between clusters they are independent, hence, sites which are in the border of a cluster are not dependent on a site which belongs to another cluster even though they are neighbors. The CAR model fails to address this property of spatially clustered data. It assumes every site is dependent on its neighboring sites although they are from different clusters. For this reason, our proposed GEE based model with a scaled-distance based correlation structure preforms better than CAR both in terms of parameter estimation and prediction performance of spatial effect. The relative performance of the proposed model is increasingly better in case of stronger clustering pattern which is demonstrated in terms of different values of the scale parameter λ defined in equation (11). The reason for such relative advantage of the proposed method over CAR is that, it com-



(a) True Process



(b) Fitted Process by GEE



(c) Fitted Process by CAR

Figure 5: Comparison of the fitted vitamin A supplement coverage among postpartum women processes by the proposed GEE approach and the CAR model with the observed process. Division boundaries are marked by dark lines. District names mentioned in the discussions are marked by numbers as: 1. Habiganj, 2. Kishorganj, 3. Mymensing, 4. Brhmanbaria, 5. Gazipur, 6. Natore, 7. Kushtia, 8. Khagrachari, 9. Bandarban.

pensate the error induced in CAR model by considering both the within cluster neighbouring units and the bordering between cluster neighbouring units as equally auto-correlated in case of clustered data where there is a presence of strong within cluster spatial auto-correlation and no between cluster auto-correlation.

These findings were also reflected in the analysis of the real life data on coverage rates of vitamin A supplement for postpartum women in Bangladesh. It is, therefore, recommended that care and attention should be paid in choice of model for producing spatially revised estimates of characteristics with known clustered pattern, a GEE approach with the suggested correlation structure will perform better than the usual CAR model in such situations. In particular, when the characteristic of interest is expected outcome of a sub-regionally administered program or intervention, such differences become crucial in monitoring and evaluation of the program or intervention.

References

Albert PS, McShane LM (1995). "A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data." *Biometrics*, **51**, 627–638.

Augustin NH, Kublin E, Metzler B, Meierjohann E, von Wühlisch G (2005). "Analyzing the Spread of Beech Canker." *Forest Science*, **51**(5), 438–448.

Banerjee S, Carlin BP, Gelfand AE (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Crc Press.

Besag J (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236.

Carl G, Kühn I (2007). "Analyzing Spatial Autocorrelation in Species Distributions Using Gaussian and Logit Models." *Ecological Modelling*, **207**(2), 159–170.

Cressie N (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Dormann CF (2007). "Assessing the Validity of Autologistic Regression." *Ecological Modelling*, **207**(2), 234–242.

EPI Coverage Evaluation Survey (2009). "Expanded Programme on Immunization." *DGHS, Mohakhali, Dhaka 1212*.

F Dormann C, M McPherson J, B Araújo M, Bivand R, Bolliger J, Carl G, G Davies R, Hirzel A, Jetz W, Daniel Kissling W, *et al.* (2007). "Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review." *Ecography*, **30**(5), 609–628.

Gotway CA, Stroup WW (1997). "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction." *Journal of Agricultural, Biological, and Environmental Statistics*, pp. 157–178.

Gumpertz L, Pye M, *et al.* (2000). "Logistic Regression for Southern Pine Beetle Outbreaks with Spatial and Temporal Autocorrelation." *Forest Science*, **46**(1), 95–107.

Hin LY, Wang YG (2009). "Working-correlation-structure Identification in Generalized Estimating Equations." *Statistics in Medicine*, **28**(4), 642–658.

Jacquez GM (2008). "Spatial Cluster Analysis." *The Handbook of Geographic Information Science*, **395**, 416.

Knox E (1989). "Detection of Clusters." *Methodology of Enquiries into Disease Clustering. London: Small Area Health Statistics Unit*, **17**, 20.

Lawson AB, Denison DGT (2002). *Spatial Cluster Modelling*. CRC press.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**(1), 13–22.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Affiliation:

Nasrin Lipi
Institue of Statistical Research and Training
University of Dhaka
Dhaka- 1000, Bangladesh
E-mail: nlipi@isrt.ac.bd

Mohammad Samsul Alam
Institue of Statistical Research and Training
University of Dhaka
Dhaka- 1000, Bangladesh
E-mail: msalam@isrt.ac.bd

Syed Shahadat Hossain
Institue of Statistical Research and Training
University of Dhaka
Dhaka- 1000, Bangladesh
E-mail: shadahat@isrt.ac.bd

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2020-01-22
Accepted: 2020-04-28