Maximum Likelihood Estimation for Directional Conditionally Autoregressive Models

Minjung Kyung * and Sujit K. Ghosh *

Institute of Statistics Mimeo Series #2597

Abstract

A spatial process observed over a lattice or a set of irregular regions is usually modeled using a conditionally autoregressive (CAR) model. The neighborhoods within a CAR model are generally formed using only the inter-distances between the regions. To accommodate the effect of directions, a new class of spatial models is developed using different weights given to neighbors in different directions. The proposed model generalizes the usual CAR model by accounting for spatial anisotropy. Maximum likelihood estimators are derived and shown to be consistent under some regularity conditions. Simulation studies are presented to evaluate the finite sample performance of the new model as compared to CAR model. Finally the method is illustrated using a data set on the crime rates of Columbus, OH.

Key words: Anisotropy; Condionally autoregressive models; Lattice data, Maximum likelihood estimation, Spatial analysis.

*Minjung Kyung is a graduate student and Sujit K. Ghosh is an associate professor, both at the Department of Statistics, North Carolina State University, Raleigh, NC, USA.

1 Introduction

In many studies, counts or averages over arbitrary regions, known as lattice or area data (Cressie, 1993), are observed and spatial analysis is performed. Given a set of geographical regions, observations collected over regions nearer to each other tend to have similar characteristics as compared to distant regions. In geoscience, this feature is known as the *Tobler's first law* (Miller, 2004). From a statistical perspective, this feature is attributed to the fact that the autocorrelation between the observations collected from nearer regions tends to be higher than those that are distant.

In general, given a set of regions $S_1, \ldots, S_n$, we consider a generalized linear model for the aggregated responses, $Y_i = Y(S_i)$, as

$$\begin{aligned} \mathrm{E}[Y_i|Z_i] &= g(Z_i) \\ \text{and } Z_i &= \mu_i + \eta_i \quad i = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where $g(\cdot)$ is a suitable link function, μ_i's denote large-scale variations and η_i's represent small-scale variations (or spatial random effects). The latent spatial process Z_i's are usually modeled using a conditionally autoregressive (CAR) model (Besag, 1974) or a simultaneously autoregressive (SAR) model (Ord, 1975). These models have been widely used in spatial statistics (Cliff and Ord, 1981). The CAR and SAR models are used to study how a particular region is influenced by its "neighboring regions". The large sacle variations, μ_i's are ususally modeled as a function of some predictor variables (e.g., latitudes, longitudes and other areal level covariates) using a parametric or semiparametric regression model (see van der Linde et al., 1995). In this article, we instead focus on developing more flexible models for the spatial random effects η_i's.

Gaussian CAR models have been used as random effects within generalized mixed

effects models to explain the latent spatial process using suitably formed neighbors (Breslow and Clayton, 1993). Gaussian CAR process has the merit that the finite dimensional joint distributions of the spatial process are multivariate Gaussian distributions. So, the maximum likelihood (ML) and the Bayesian estimates are easily obtained. Howver one of the major limitations of the CAR model is that the neighbors are formed using some form a distance metric and the effect of the direction is completely ignored. However, if the underlying spatial process is anisotropic, the magnitude of autocorrelation between the neighbors might be different in different directions. This limitation serves as our main motivation and an extension of the regular CAR process is proposed that can capture anisotropy. The newly proposed spatial process will be termed as the directional CAR (DCAR) model. In Section 2, we define the new spatial process and present statistical inferences for the parameters based on maximum likelihood (ML) theory. In Section 2.2, the ML estimator is shown to be consistent under some regularity conditions. Section 3 presents the finite sample performance of the ML estimators and the newly proposed DCAR models are compared against the regular CAR models in terms of popular information theoretic criteria. In Section 4, the proposed method is illustrated and compared with regular CAR using a data set of the crime rates in Columbus, OH. Finally, in Section 5, some extensions of the DCAR model are discussed.

2 Directional CAR models

Consider again the model described by (1). In this section, we develop a new model for the Z_i's. For illustrations and notation simplicity assume that S_i are regions in a two-dimensional space, i.e., $S_i \in \mathbb{R}^2,\ \forall i$. However, the model and associated statistical

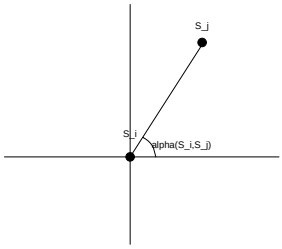

Figure 1: The angle (in radian) α_{ij}

inference presented in this article can easily be extended to higher dimensional data. Let $s_i = (s_{1i}, s_{2i})$ be a centroid of the sub-region S_i, where s_{1i} corresponds to the horizontal coordinate (x-coordinate) and s_{2i} corresponds to the vertical coordinate (y-coordinate). The angle (in radian) between s_i and s_j is defined as

$$\alpha_{ij} = \alpha(S_i, S_j) = \begin{cases} \left| \tan^{-1}(\frac{s_{2j}-s_{2i}}{s_{1j}-s_{1i}}) \right| & \text{if } s_{2j} - s_{2i} \geq 0 \\ -\left(\pi - \left| \tan^{-1}(\frac{s_{2j}-s_{2i}}{s_{1j}-s_{1i}}) \right|\right) & \text{if } s_{2j} - s_{2i} < 0 \end{cases}$$

for all $j \neq i$. We consider directions of neighbors from the centroid of subregion S_i's. For example, in Figure 1, S_j is in the north-east region of S_i and hence $\alpha(S_i, S_j)$ is in $[0, \frac{\pi}{2})$. Let $\mathcal{N}_i$ represents a set of indexes of neighborhoods for the i-th region S_i that are based on some form distance metric (say as in a regular CAR model). We can now create new sub-regions, for each i, as follows:

$$\begin{aligned} \mathcal{N}_{i1} &= \{j : j \in \mathcal{N}_i, 0 \leq \alpha_{ij} < \frac{\pi}{2}\}, \\ \mathcal{N}_{i2} &= \{j : j \in \mathcal{N}_i, \frac{\pi}{2} \leq \alpha_{ij} < \pi\}, \\ \mathcal{N}_{i3} &= \{j : j \in \mathcal{N}_i, \pi \leq \alpha_{ij} < \frac{3}{2}\pi\}, \\ \mathcal{N}_{i4} &= \{j : j \in \mathcal{N}_i, \frac{3}{2}\pi \leq \alpha_{ij} < 2\pi\}. \end{aligned}$$

However, these directional neighborhoods should be chosen carefully so that for each i they form a *clique*. Recall that a *clique* is any set of sites which either consists of

a single site or else in which every site is a neighbor of every other site in the set (Besag, 1974).

This would allow us to show the existence of the spatial process by using the Hammersley-Clifford Theorem (Besag, 1974 p.197-198) and to derive the finite dimensional joint distribution of process using only a set of full conditional distributions. For instance, if $j \in \mathcal{N}_{i1}$ then we should ensure that $i \in \mathcal{N}_{j3}$. For the above set of four sub-neighborhoods, we can combine each pair of the diagonally opposite neighborhoods to form a new neighborhood, i.e., we can create $\mathcal{N}^*_{i1} = \mathcal{N}_{i1} \bigcup \mathcal{N}_{i3}$, and $\mathcal{N}^*_{i2} = \mathcal{N}_{i2} \bigcup \mathcal{N}_{i4}$ for $i = 1, \ldots, n$. Now it is easy to check that if $j \in \mathcal{N}^*_{i1}$, then $i \in \mathcal{N}^*_{j1}$. Thus, we redefine two subsets of $\mathcal{N}_i$'s as follows:

$$\begin{aligned} \mathcal{N}^*_{i1} &= \{j : j \in \mathcal{N}_i \text{ and } (0 \leq \alpha_{ij} < \frac{\pi}{2} \text{ or } \pi \leq \alpha_{ij} < \frac{3}{2}\pi)\} \\ \mathcal{N}^*_{i2} &= \{j : j \in \mathcal{N}_i \text{ and } (\frac{\pi}{2} \leq \alpha_{ij} < \pi \text{ or } \frac{3}{2}\pi \leq \alpha_{ij} < 2\pi)\}. \end{aligned} \tag{2}$$

Then, each of $\mathcal{N}^*_{i1}$ and $\mathcal{N}^*_{i2}$ forms a clique and that $\mathcal{N}_i = \mathcal{N}^*_{i1} \bigcup \mathcal{N}^*_{i2}$. The above scheme of creating new neighborhoods based on the inter-angles, α_{ij}'s can be extended beyond just two sub-neighborhoods so that each of the new sub-neighborhood forms a clique. But for the rest of the article we restrict our attention to case with only two sub-neighborhoods as described in (2).

Based on subsets of the associated neighborhoods, $\mathcal{N}^*_{i1}$ and $\mathcal{N}^*_{i2}$, we can construct directional weight matrices $\mathbf{W}^{(1)} = ((w^{(1)}_{ij}))$ and $\mathbf{W}^{(2)} = ((w^{(2)}_{ij}))$, respectively. For instance, we can define the directional proximity matrices as $w^{(1)}_{ij} = 1$ if $j \in \mathcal{N}^*_{i1}$ and $w^{(2)}_{ij} = 1$ if $j \in \mathcal{N}^*_{i2}$. Notice that $\mathbf{W} = \mathbf{W}^{(1)} + \mathbf{W}^{(2)}$ reproduces the commonly used proximity matrix based on distances as in a regular CAR model.

In order to model the large-scale variations, without loss of any generality we assume a canonical linear model, $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$, where $\mathbf{x}_i$'s are vectors of predictor variables

and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$ is a vector of regression coefficients. Notice that nonlinear regression functions involving smoothing splines and polynomials can be re-written in the above canonical form (e.g., see Wahba, 1977). From model (1) it follows that

$$\mathrm{E}[\boldsymbol{Z}] = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Var}[\boldsymbol{Z}] = \boldsymbol{\Sigma}(\boldsymbol{\omega}), \tag{3}$$

where $\boldsymbol{\omega}$ denotes the vector of spatial autocorrelation parameters and other variance components. Notice that along with (3) the model (1) can be used for discrete responses using a generalized linear model framework (Schabenberger and Gotway, 2005, p.353). We now develop a model for $\boldsymbol{\Sigma}(\boldsymbol{\omega})$ that accounts for anisotropy.

Let δ_1 and δ_2 denote the directional spatial effects corresponding to $\mathcal{N}_{i1}$'s and $\mathcal{N}_{i2}$'s, respectively. We define the distribution of Z_i conditional on the rest of Z_j's for $j \neq i$ using only the first two moments:

$$\begin{aligned} \mathrm{E}[Z_i | Z_j = z_j j \neq i, \mathbf{x}_i] &= \mathbf{x}_i^T\boldsymbol{\beta} + \sum_{k=1}^{2} \delta_k \sum_{j=1}^{n} w_{ij}^{(k)} \big(z_j - \mathbf{x}_j^T\boldsymbol{\beta}\big) \\ \mathrm{Var}[Z_i | Z_j = z_j j \neq i, \mathbf{x}_i] &= \frac{\sigma^2}{m_i}, \end{aligned} \tag{4}$$

where $w_{ij}^{(k)} \geq 0$ and $w_{ii}^{(k)} = 0$ for $k = 1, 2$ and $m_i = \sum_{j=1}^{n} w_{ij}$.

The joint distribution based on a given set of full conditional distributions can be derived using the *Brook's Lemma* (Brook, 1964), provided the positivity condition is satisfied. For the DCAR model, by construction it follows that each of $\mathcal{N}_{i1}^*$ and $\mathcal{N}_{i2}^*$ defined in (2) forms a clique for $i = 1, \ldots, n$. Thus, it follows from the Hammersley-Clifford Theorem that the latent spatial process Z_i of a DCAR model exists and is a Markov Random Field (MRF). We now derive the exact joint distribution of the Z_i's by assuming that each of the full conditional distribution is a Gaussian distribution.

2.1 Gaussian DCAR models

The Gaussian CAR model has been used widely for the latent spatial process Z_i. In this section, we study merits of the Gaussian DCAR model. As we discussed in previous section, the joint distribution can be easily derived from the conditional distributions by using Brook's Lemma.

Assume that the full conditional distributions of Z_i's are given as

$$Z_i|Z_j = z_j, j \neq i, \mathbf{x}_i \sim N\bigg(\mathbf{x}_i^T\boldsymbol{\beta} + \sum_{k=1}^{2}\delta_k \sum_{j=1}^{n} w_{ij}^{(k)}\big(z_j - \mathbf{x}_j^T\boldsymbol{\beta}\big), \frac{\sigma^2}{m_i}\bigg), \tag{5}$$

where $w_{ij}^{(k)}$ for $k = 1, 2$ are the directional weights. It can be shown that this latent spatial DCAR process Z_i's is a MRF. Thus, by Brook's lemma and the Hammersley-Clifford Theorem, it follows that the finite dimensional joint distribution is a multivariate Gaussian distribution given by

$$\boldsymbol{Z} \sim N_n\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} - \delta_1\mathbf{W}^{(1)} - \delta_2\mathbf{W}^{(2)})^{-1}\mathbf{D}\big), \tag{6}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$ and $\mathbf{D} = \text{diag}(\frac{1}{m_1}, \ldots, \frac{1}{m_n})$. For simplicity, we denote the variance-covariance matrix of DCAR process by $\boldsymbol{\Sigma}_Z^* \equiv \sigma^2(\mathbf{I} - \delta_1\mathbf{W}^{(1)} - \delta_2\mathbf{W}^{(2)})^{-1}\mathbf{D}$.

For a proper Gaussian model, the variance-covariance matrix $\boldsymbol{\Sigma}_Z^*$ is required to be positive definite. If we use the standardized directional proximity matrices $\tilde{\mathbf{W}}^{(k)} = ((\tilde{w}_{ij}^{(k)} = \frac{w_{ij}^{(k)}}{m_i}))$, $k = 1, 2$, it can be easily shown that $\boldsymbol{\Sigma}_Z^*$ is symmetric.

Next, we derive a sufficient condition that ensures that the variance-covariance matrix $\boldsymbol{\Sigma}_Z^*$ is non-singular. As $\mathbf{D}$ is a diagonal matrix, we only require suitable conditions on $\tilde{\mathbf{W}}^{(k)}$ and on δ_k for $k = 1, 2$. The following results provides a sufficient condition:

Lemma 1 *Let* $\mathbf{A} = \mathbf{I} - \sum_{k=1}^{K}\delta_k\tilde{\mathbf{W}}^{(k)}$ *be an* $n \times n$ *matrix where* $\sum_{k=1}^{K}\tilde{\mathbf{W}}^{(k)}$ *is a*

symmetric matrix with non-negative entries, diagonal $\mathbf{0}$ *and each row sums to unity. If* $\max_{1\le k\le K}|\delta_k| < 1$*, then the matrix* $\mathbf{A}$ *is positive definite.*

Proof: Let a_{ij} denote the (i,j)-th element of $\mathbf{A}$. Notice that for each $i = 1, 2\ldots, n$, we have

$$\sum_{j\neq i}|a_{ij}| = \sum_{j\neq i}|\sum_{k=1}^{K}\delta_k w_{ij}^{(k)}| \le \sum_{k=1}^{K}|\delta_k|\sum_{j\neq i}w_{ij}^{(k)} < \sum_{k=1}^{K}\sum_{j\neq i}w_{ij}^{(k)} = 1 = a_{ii}$$

Hence it follows from Lemma 2 (see Appendix A) that $\mathbf{A}$ is positive definite.

Notice that when $\delta_1 = \delta_2 = \rho$, DCAR$(\delta_1, \delta_2, \sigma^2)$ reduces to CAR(ρ, σ^2). The next step of a statistical analysis is to estimate unknown parameters of the DCAR model based on the observed responses and the explanatory variables and to predict future values at unobserved sites. Thus, in the next section, we discuss the maximum likelihood (ML) methods for the spatial autoregressive models.

2.2 Parameter estimation using ML theory

With Gaussian DCAR model of the latent spatial process Z_i's, we describe how to estimate parameters and associated measures of uncertainties based on ML methods.

Mardia and Marshall (1984) have given sufficient conditions in the case of Gaussian process for the consistency and asymptotic normality of the ML estimators. Based on their results, first we describe the numerical computation of the MLE of $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \sigma^2, \boldsymbol{\delta}^T)^T$, where $\boldsymbol{\delta} = (\delta_1, \delta_2)^T$, for a Gaussian DCAR model of Z_i's. Then, we show that the ML estimator of Gaussian DCAR model is consistent and asymptotically normal.

In the case of autoregressive models, likelihood is a complex nonlinear function of the spatial correlation parameters. Thus, the log-likelihood of Gaussian DCAR model

is a complex nonlinear function of δ_1 and δ_2 which appear in the variance-covariance matrix. Here, for the latent spatial DCAR process Z_i's, under the joint multivariate Gaussian distribution, the log-likelihood function is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln|\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}| \\ &\quad -\frac{1}{2\sigma^2}(\boldsymbol{Z} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{D}^{-1}\mathbf{A}^*(\boldsymbol{\delta})(\boldsymbol{Z} - \mathbf{X}\boldsymbol{\beta}) \qquad (7)\end{aligned}$$

where $\mathbf{A}^*(\boldsymbol{\delta}) = \mathbf{I} - \delta_1\tilde{\mathbf{W}}^{(1)} - \delta_2\tilde{\mathbf{W}}^{(2)}$ and $\mathbf{D} = \text{diag}(\frac{1}{m_1}, \ldots, \frac{1}{m_n})$. Thus, for known $\boldsymbol{\delta}$, the profile ML estimators of $\boldsymbol{\beta}$ and σ^2 are given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}(\boldsymbol{\delta})} &= \left(\mathbf{X}^T\mathbf{D}^{-1}\mathbf{A}^*(\boldsymbol{\delta})\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}^{-1}\mathbf{A}^*(\boldsymbol{\delta})\boldsymbol{Z} \\ \widehat{\sigma^2(\boldsymbol{\delta})} &= n^{-1}(\boldsymbol{Z} - \mathbf{X}\widehat{\boldsymbol{\beta}(\boldsymbol{\delta})})^T\mathbf{D}^{-1}\mathbf{A}^*(\boldsymbol{\delta})(\boldsymbol{Z} - \mathbf{X}\widehat{\boldsymbol{\beta}(\boldsymbol{\delta})}).\end{aligned}$$

Substituting back into (7), the ML estimator of $\boldsymbol{\delta}$ can be obtained by maximizing the profile log likelihood,

$$\ell^*(\boldsymbol{\delta}) = -\frac{n}{2}\ln(\widehat{\sigma^2(\boldsymbol{\delta})}) + \frac{1}{2}\ln|\mathbf{A}^*(\boldsymbol{\delta})| \qquad (8)$$

which can be numerically maximized to obtain $\widehat{\delta_1}$ and $\widehat{\delta_2}$.

In order to maintain the restriction $\max(|\delta_1|, |\delta_2|) < 1$, we use the "L-BFGS-B" method (Byrd, et al., 1995) to compute the ML estimates maximizing the log-likelihood. We also extract the Hessian matrix to estimate the information matrix. In other words, we use observed Fisher information matrix to obtain the standard errors (s.e.'s) of $\boldsymbol{\eta}$. It has been shown by Efron and Hinkley (1978) that use of the observed Fisher information matrix results into more efficient estimator than the use of the expected Fisher information matrix.

For large samples, the sampling distribution of the MLE will be shown to have an asymptotically normal distribution centered at the true value of parameters with

variance determined by the Fisher information. To derive the Fisher information matrix of MLE's for Gaussian DCAR model, the steps of p. 483-484 in Section 7.3.1 of Cressie (1993) are followed. However, for the derivatives of the variance-covariance matrix, Lemma 3 and Lemma 4 in Appendix A are used. For a DCAR model, we define a vector of variance-covariance parameters as $\boldsymbol{\gamma} = (\sigma^2, \boldsymbol{\delta}^T)^T = (\gamma_1 = \sigma^2, \gamma_2 = \delta_1, \gamma_3 = \delta_2)^T$. Then, we denote $\boldsymbol{\Sigma}_{(i)} = \partial\boldsymbol{\Sigma}/\partial\boldsymbol{\gamma}_i$ and $\boldsymbol{\Sigma}^{(i)} = \partial\boldsymbol{\Sigma}^{-1}/\partial\boldsymbol{\gamma}_i = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}^{-1}$ for $i = 1, 2, 3$. Also, $\boldsymbol{\Sigma}_{(ij)} = \partial^2\boldsymbol{\Sigma}/\partial\gamma_i\partial\gamma_j$ and $\boldsymbol{\Sigma}^{(ij)} = \partial^2\boldsymbol{\Sigma}^{-1}/\partial\gamma_i\partial\gamma_j$ for $i, j = 1, 2, 3$. Thus, for a DCAR model,

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{(1)} = -(\sigma^{-2})\mathbf{I}_{n\times n}, \quad \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{(2)} = -\mathbf{A}^*(\boldsymbol{\delta})^{-1}\tilde{\mathbf{W}}^{(1)} \quad \text{and } \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{(3)} = -\mathbf{A}^*(\boldsymbol{\delta})^{-1}\tilde{\mathbf{W}}^{(2)}.$$

Therefore, the Fisher information matrix for Gaussian DCAR model has the form

$$\mathbf{I}(\boldsymbol{\eta}) = \begin{bmatrix} \sigma^{-2}\mathbf{X}^T\mathbf{D}^{-1/2}\mathbf{A}^*(\boldsymbol{\delta})\mathbf{D}^{-1/2}\mathbf{X} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}^T & \frac{n}{2}\sigma^{-4} & \frac{1}{2}\sigma^{-2}tr(\mathbf{G}_1) & \frac{1}{2}\sigma^{-2}tr(\mathbf{G}_2) \\ \mathbf{0}^T & \frac{1}{2}\sigma^{-2}tr(\mathbf{G}_1) & \frac{1}{2}\nu_1 & \frac{1}{2}tr(\mathbf{G}_1\mathbf{G}_2) \\ \mathbf{0}^T & \frac{1}{2}\sigma^{-2}tr(\mathbf{G}_2) & \frac{1}{2}tr(\mathbf{G}_1\mathbf{G}_2) & \frac{1}{2}\nu_2 \end{bmatrix}, \tag{9}$$

where $\mathbf{G}_k = \tilde{\mathbf{W}}^{(k)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}$, $\nu_k = tr(\mathbf{G}_k\mathbf{G}_k) = tr(\mathbf{G}_k^2)$, for $k = 1, 2$. We estimate this Fisher information (FI) matrix (9) to obtain the asymptotic variances of parameters using the observed FI.

Because $\boldsymbol{Z}$ is a single observation from $N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_Z)$, it is not obvious that $\widehat{\boldsymbol{\eta}}$ is consistent and asymptotically normal. Sweeting (1980) has given a general result of weak consistency and uniform asymptotic normality for MLE's based on dependent observations. Using Sweeting's result, Mardia and Marshall (1984) prove the following theorem.

Theorem 1 *(Mardia and Marshall, 1984) Suppose* $\boldsymbol{Z} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$*, where* $\boldsymbol{\beta}$ *is a*

$p \times 1$ vector of unknown large-scale parameters and $\boldsymbol{\Sigma}$ is function of $\boldsymbol{\gamma}$, a $k \times 1$ vector of unknown small-scale (spatial) parameters. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of $\boldsymbol{\Sigma}$ and let those of $\boldsymbol{\Sigma}_{(i)}$ and $\boldsymbol{\Sigma}_{(ij)}$ be $\{\lambda_l^i : l = 1, \ldots, n\}$ and $\{\lambda_l^{ij} : l = 1, \ldots, n\}$, with $|\lambda_1^i| \leq \cdots \leq |\lambda_n^i|$ and $|\lambda_1^{ij}| \leq \cdots \leq |\lambda_n^{ij}|$ for $i, j = 1, \ldots, k$. Suppose that, as $n \to \infty$,

(i) $\lim \lambda_n = C < \infty$, $\lim |\lambda_n^i| = C_i < \infty$, $\lim |\lambda_n^{ij}| = C_{ij} < \infty$ *for all* $i, j = 1, \ldots, k$

(ii) $||\boldsymbol{\Sigma}_{(i)}||^{-2} = O(n^{-\frac{1}{2}-\alpha})$, *for some* $\alpha > 0$ *for* $i = 1, \ldots, k$ *($||\mathbf{G}||$ denotes the Euclidean matrix norm,* $(\sum_{i,j} g_{ij}^2)^{1/2} = \{tr(G^T G)\}^{1/2}$*)*

(iii) for all $i, j = 1, \ldots, k$, $a_{ij} = \lim\{t_{ij}/(t_{ii}t_{jj})^{\frac{1}{2}}\}$ *exists, where* $t_{ij} = tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(j)})$ *and* $\mathbf{A} = ((a_{ij}))$ *is a nonsingular matrix*

(iv) $\lim(\mathbf{X}^T\mathbf{X})^{-1} = 0$.

Then these conditions are sufficient for the asymptotic normality and weak consistency of $\widehat{\boldsymbol{\eta}}$; that is $\widehat{\boldsymbol{\eta}} \sim N(\boldsymbol{\eta}, \mathbf{I}(\boldsymbol{\eta}))$, where $\mathbf{I}(\boldsymbol{\eta})$ is the information matrix and $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$.

For a proof see the paper of Mardia and Marshall (1984).

We show that the above sufficient conditions are satisfied by the MLE of the DCAR model under some additional regularity conditions and hence we establish the asymptotic normality and weak consistency of $\widehat{\boldsymbol{\eta}}$ under DCAR model.

Theorem 2 *The conditions stated in Theorem 1 are satisfied by the DCAR model if* $0 < \liminf m_{(1)} \leq \limsup m_{(n)} < \infty$ *as* $n \to \infty$ *where m_i is number of neighbors of subregion S_i,* $m_{(1)} = \min(m_1, \ldots, m_n)$ *and* $m_{(n)} = \max(m_1, \ldots, m_n)$.

The proof of the Theorem 2 is given in Appendix B.

In order to study the finite sample performance of ML estimators, we conduct a simulation study. In this simulation study, we focus on the behavior of Gaussian DCAR model of the latent spatial process $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ as defined in (6).

2.3 A simulation study

Mardia and Marshall (1984) conducted a simulation study with 10×10 unit spacing lattice, based on samples generated from normal distribution with mean zero and a spherical covariance model. The sampling distribution of the MLE's of the parameters were studied based on 300 Monte Carlo samples. Following a similar setup, for our simulation study, we selected an 15×15 unit spacing lattice and generated 500 samples (N=500) Z_i's from a multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and the variance-covariance $\sigma^2 \mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}$, where $\mathbf{A}^*(\boldsymbol{\delta}) = \mathbf{I} - \delta_1 \tilde{\mathbf{W}}^{(1)} - \delta_2 \tilde{\mathbf{W}}^{(2)}$. The $\mathbf{X}$ matrix was chosen to represent the lattice nad longitude coordinates in addition to an intercept. The true value of the parameters were fixed at $\boldsymbol{\beta} = (1, -1, 2)^T$ and $\sigma^2 = 2$. For the above mentioned DCAR model, to study the sampling distributions of MLE's $\widehat{\delta_1}$ and $\widehat{\delta_2}$, we consider four different sets of $\boldsymbol{\delta}$'s:

Case 1: $\delta_1 = -0.95$ & $\delta_2 = -0.97$: negative boundary points

Case 2: $\delta_1 = -0.30$ & $\delta_2 = 0.95$: negative point and positive boundary point

Case 3: $\delta_1 = -0.95$ & $\delta_2 = 0.97$: negative boundary point and positive boundary point

Case 4: $\delta_1 = 0.95$ & $\delta_2 = 0.93$: positive boundary points.

It is expected that for the values of the parameters near the boundary of the parameter space, there might be unexpected behaviors of the MLE's.

As we discussed in Section 2.2, in order to maintain the restriction $\max(|\delta_1|, |\delta_2|) < 1$, we use the "L-BFGS-B" method to compute the ML estimates within the `optim` function of the software `R` to maximize the log-likelihood. To obtain the standard errors (s.e.'s) of $\boldsymbol{\eta}$, it is also computationally easier to directly obtain the Hessian matrix from the output of the `optim` function.

We study the sampling distribution of parameters numerically by using tables.

First, from the Table 1, notice that there are few missing estimates of the estimated standard errors (ESE) based on the observed Fisher information matrix. This is due to the fact that `R` has a function `deriv` that numerically computes the derivative of complicated expressions which can lead to numerical instabilities. We observe that for all the four choices of $\boldsymbol{\delta}$, there are no significant biases (all p-values are bigger than 0.42). Furthermore, it appears that the estimated standard errors (ESE's) of $\widehat{\delta_1}$ and $\widehat{\delta_2}$ are good approximations to finite sample Monte Carlo standard errors (MCSE's) when the true values are positive. For all these boundary values of $\boldsymbol{\delta}$, the nominal 95% coverage probability (CP) is relatively away from 0.95 indicating that the MLE tends to estimate the true value with higher uncertainty when all δ_1 and δ_2 are near the boundary. The high coverage probability of MLE based on a symmetric confidence interval may be due to the skewness of the sampling distribution that we have observed in our empirical studies. It was observed that the estimates appear to be skewed to right for the negative extreme value and they are skewed to the left for the positive extreme value.

For the results of ML estimation of σ^2 of DCAR model, the bias of σ^2 tends to be slightly negative. In other words, the MLE's of σ^2 tend to underestimate the true values. However, it is to be noted that these biases are not statistically significant (all four p-values are greater than 0.6). Thus, for all cases, MLE of σ^2 is reasonable estimate.

For the estimation of $\boldsymbol{\beta}$'s, we observe that the finite sample performance of MLE's of $\boldsymbol{\beta}$'s of DCAR model are close to the large sample Gaussian approximation. Also, biases are not significant and nominal coverage probabilities were all found to be close to 0.95 (results not reported due to lack of space).

Table 1: Performance of MLE's of δ_1's, δ_2's and σ^2's

	δ_1	δ_2	σ^2	δ_1	δ_2	σ^2
True	-0.95	-0.97	2.00	-0.30	0.95	2.00
bias	0.13	0.21	-0.09	-0.19	-0.25	-0.17
MCSE	0.26	0.39	0.34	0.37	0.31	0.42
P-value	0.62	0.60	0.80	0.60	0.42	0.68
ESE	0.46	0.47	0.19	0.38	0.38	0.17
	(N=493)	(N=494)		(N=491)	(N=491)	
95% CP	0.98	0.96	0.90	0.88	0.95	0.83
	(N=493)	(N=494)		(N=491)	(N=491)	

	δ_1	δ_2	σ^2	δ_1	δ_2	σ^2
True	-0.95	0.97	2.00	0.95	0.93	2.00
bias	0.03	-0.17	-0.09	-0.16	-0.14	-0.02
MCSE	0.16	0.25	0.35	0.24	0.23	0.27
P-value	0.86	0.50	0.80	0.50	0.54	0.94
ESE	0.33	0.34	0.18	0.28	0.28	0.19
	(N=487)	(N=492)		(N=494)	(N=494)	(N=499)
95% CP	0.99	0.97	0.84	0.98	0.98	0.94
	(N=487)	(N=492)		(N=494)	(N=494)	(N=499)

3 Model comparison using information criteria

The DCAR model is a generalization of the CAR model. So it would be interesting to investigate what happens when someone fits a CAR model to a data set generated under a DCAR model and vice versa. To compare models, we consider the Akaike's Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwartz, 1978) based on the ML method. AIC, a penalized-log-likelihood criterion, is defined by

$$AIC = -2\ell(\widehat{\boldsymbol{\eta}}) + 2k,$$

where $\ell(\boldsymbol{\eta})$ is defined in (7) and k is number of parameters. Also, BIC is defined by

$$BIC = -2\ell(\widehat{\boldsymbol{\eta}}) + k\ln(n),$$

where n is the number of observations. In theory, we select the model with the smaller AIC and BIC values. We consider two cases based on data generated from a (i) CAR model and (ii) DCAR model.

3.1 Results based on CAR data

With samples from Gaussian CAR process, we fit both CAR and DCAR models, respectively. Notice that if there is no directional difference in the observed spatial data, then the estimates of δ_1 should be very similar to the estimates of δ_2. Thus we might expect very similar estimates for δ_1 and δ_2 based on a sample from CAR process because $CAR(\rho, \sigma^2) = DCAR(\rho, \rho, \sigma^2)$. To study the sampling distribution of ρ, we consider five different values of ρ's, Case 1:$\rho = -0.95$, Case 2:$\rho = -0.25$, Case 3:$\rho = 0$, Case 4:$\rho = 0.25$ and Case 5:$\rho = 0.95$. For each case, we compare the estimates of CAR and DCAR model with ML method.

From the results based on the ML estimates of CAR model, we observe that for all the cases there are no significant biases at 5% level. The ESE of ρ is a good approximation to finite sample variance when the spatial dependence is weak such as in cases 2,3, and 4, because the ESE's are close to MCSE's. However, when the true value is near negative boundary, MLE tends to estimate the true value with high uncertainty. For all cases, the biases of MLE's of σ^2 tend to be slightly negative. However, these negative biases are not significant (all p-values are bigger than 0.7). For the estimation of $\boldsymbol{\beta}$'s, the estimates did not have any significant bias (all p-values are bigger than 0.9). We have presented these results due to lack of space but are available in the first author's doctoral thesis.

3.1.1 Estimation under model misspecification

From the results of the ML estimates of DCAR model, we observe that the ML estimates of $\boldsymbol{\delta}$ and σ^2 are slightly negatively biased for all cases except when the true value of ρ is near negative boundary, $\rho = -0.95$. However, based on p-value, we conclude that there is no significant bias (all p-values are bigger than 0.5). We notice that the ML estimates of $\boldsymbol{\delta}$ and σ^2 of DCAR model are similar to the estimates of ρ and σ^2 of CAR model, respectively. It means that with data sets from a CAR model, the sampling distributions of parameters are quite similar for either fitting CAR or DCAR model. However, the estimates of $\boldsymbol{\delta}$ for each cases have larger MCSE and ESE values than the estimates of ρ as expected. The estimates of $\boldsymbol{\beta}$ based on DCAR model is not affected at all even if the data was generated from a CAR model. This is also expected as under both models $\mathbf{X}\boldsymbol{\beta}$ is the mean of $\boldsymbol{y}$.

Instead of comparing the parameter estimates of CAR and DCAR models based on samples from CAR process, we use AIC and BIC to compare the overall model

Table 2: Compariosn of AIC and BIC between CAR and DCAR models with data sets from CAR process (PCD = Percentage of Correct Decision)

DGP	CAR($\rho = -0.95$)		CAR($\rho = -0.25$)		CAR($\rho = 0.00$)	
Fit	CAR	DCAR	CAR	DCAR	CAR	DCAR
PCD(AIC)	90%	10%	81%	19%	88%	12%
P-value	0.01		0.94		0.92	
PCD(BIC)	95%	5%	94%	6%	95%	5%
P-value	0.01		0.93		0.91	
DGP	CAR($\rho = 0.25$)		CAR($\rho = 0.95$)			
Fit	CAR	DCAR	CAR	DCAR		
PCD(AIC)	84%	16%	95%	5%		
P-value	0.63		0.51			
PCD(BIC)	96%	4%	98%	2%		
P-value	0.61		0.51			

performance. It has been shown by Stone (1977) that the use of AIC is similar to cross-validation. We define DGP as the data generating process and FIT as the data fitting models. Also, we define PDC as the percentage of correct decision. From Table 2, we observe that AIC and BIC of CAR model are smaller than those of DCAR model for the same data sets from CAR process for all five cases. PCD based on calculated AIC values is 90% for samples from CAR process when the true $\rho = -0.95$. It means that for $N = 500$ samples from CAR process with $\rho = -0.95$, if we fit both CAR and DCAR models, 90% of AIC's of CAR models is smaller than those of DCAR models. However, p-values for hypothesis that AIC or BIC values of CAR model is the same as that of DCAR model, are much larger than 0.05 for Cases 2,3,4 and 5, but for Case 1, p-values for hypothesis test are less than 0.05. In other words when the true value of ρ is near negative boundary ($\rho = -0.95$), for the data set of CAR process, CAR model fits better than DCAR model. Other cases, we can use either CAR or DCAR model for the ML method based on samples from CAR process. This is expected as CAR is nested under DCAR.

Based on the performance of AIC and BIC, we observe that for the data set generated from CAR with $\rho = -0.95$, the CAR model fits significantly better than the DCAR model. However, for all other cases that we investigated, we conclude that we may safely use DCAR model even when the data is generated from a CAR model.

3.2 Results based on DCAR data

In this section, we fit both CAR and DCAR models to the data sets generated from DCAR model. In this case, there might exist somewhat unexpected sampling distribution of ρ as it will not be able to capture the directional effects.

3.2.1 Estimation under model misspecification

From the ML estimates of ρ, it appears that $\widehat{\rho}$ seems to estimate the mean of the true values of δ_1 and δ_2 for a sample from a DCAR process. Therefore, we define the pseudo-true value of ρ as the mean of the true values of δ_1 and δ_2. By fitting DCAR and CAR model to samples from DCAR process, we observe that if the mean of δ_1 and δ_2 is near positive or negative boundary, the sampling distribution of $\widehat{\rho_0} = \frac{\widehat{\delta_1}+\widehat{\delta_2}}{2}$ is very similar to the sampling distribution of $\widehat{\rho}$. Estimates of ρ_0's are slightly negatively biased except Case 1. However, as p-values are bigger than 0.3, we conclude that there is no significant biases of all parameters (Results are not tabulated due to lack of space). For the estimates of σ^2, the estimates were found to be slightly negatively biased if δ_1 and δ_2 were near the same negative or positive boundary. However, if δ_1 and δ_2 are in negative and positive area separately, the estimates of σ^2 are slightly positively biased. Again such biases are not statistically significant because all p-values are bigger than 0.6. We notices that if the true value $\rho_0 = \frac{\delta_1+\delta_2}{2}$ is near negative boundary, the estimates of $\widehat{\rho}$ are skewed right and if ρ_0 is near positive

boundary, the estimates of $\widehat{\rho}$ are skewed left.

From the Table 3, we observe that CAR model has smaller AIC values than DCAR model for samples from DCAR process when $\delta_1 \approx \delta_2$ as expected as CAR is nested under DCAR. However, when δ_1 and δ_2 are in negative and positive valued, respectively, PCD of AIC for DCAR model is 70% for Case 2 and 3, respectively, thus AIC picks up the true model, DCAR model, more frequently. Nevertheless, p-values for hypothesis that AIC of CAR model is the same as that of DCAR model, are greater than 0.05 for all cases. From the Table 3, we observe that BIC picks up a DCAR model more frequently with percentage of correct decisions by 62%, when the data are from DCAR process with $\delta_1 = -0.95$ and $\delta_2 = 0.97$, two extreme values in each negative and positive parameter space and by 64% when the data are from DCAR process with $\delta_1 = -0.30$ and $\delta_2 = 0.95$.

Notice that BIC penalizes the DCAR model in favor of CAR model when $\delta_1 \approx \delta_2$ as expected because a DCAR with $\delta_1 = \delta_2$ reduces to a CAR model. In summary, we would recommend the use of BIC as a good model selection criteria compared to AIC in choosing the better model when comparing CAR and DCAR model.

Therefore, the CAR model captures the mean of directional spatial effects when the data is generated from a DCAR model. The use of information criterion suggests that DCAR model works better if there exist directionally different relationship within neighbors and there is no significant loss of efficiency even if the data arise from a CAR process.

Table 3: Compariosn of AIC and BIC between CAR and DCAR models with data sets from DCAR process (PCD = percentage of correct decisions)

DGP	DCAR($\delta_1 = -0.95$, $\delta_2 = -0.97$)		DCAR($\delta_1 = -0.30$, $\delta_2 = 0.95$)	
Fit	CAR	DCAR	CAR	DCAR
PCD(AIC)	91%	9%	30%	70%
P-value		0.07		0.09
PCD(BIC)	95%	5%	36%	64%
P-value		0.06		0.08
DGP	DCAR($\delta_1 = -0.95$, $\delta_2 = 0.97$)		DCAR($\delta_1 = 0.95$, $\delta_2 = 0.93$)	
Fit	CAR	DCAR	CAR	DCAR
PCD(AIC)	30%	70%	96%	4%
P-value		0.40		0.17
PCD(BIC)	38%	62%	98%	2%
P-value		0.38		0.17

4 Data analysis

We illustrate the fitting of DCAR and CAR model using a real data set by estimating the crime distribution in Columbus, Ohio collected in the year of 1980. We also use income level and housing values as predictors for crime rates. These observations were collected in 49 contiguous Planning Neighborhoods of Columbus, Ohio. Neighborhoods correspond to census tracts, or aggregates of a small number of census tracts. Original data set can be found in Table 12.1 of Anselin (1988, p.189). In this data set, the crime variable pertains to the total of residential burglaries and vehicle thefts per thousand households. The income and housing values are measured in thousands of dollars. Anselin (1988) illustrated the existence of spatial dependence by using diagnostic tests based on ordinary least squares (OLS) estimation, ML estimation using a mixed regressive-autoregressive model and ML estimation using SAR model. Notably, Anselin considered two sets of separate regression for the east and west side of the city using the SAR model. Based on the Wald, Likelihood Ratio (LR) and

Lagrange Multiplier statistics, he concluded that when SAR model is used, there exists structural instability. It means that given a SAR spatial structure, the regression equation for the east side of the city is different from that of the west. Instead of using a generalized linear model fro count data, we make a variance stabilizing log-transformation for Poisson counts and treat the crime rate to be continuous variable. With log-transformed crime rate, we assume Gaussian distribution with CAR and DCAR spatial structure.

We plot the the log-transformed crime rates that are divided into 5 intervals of each 20% quantiles in Figure 2. From original data, we observe that Y_4 and Y_{17} have extremely small values. The region S_4 is on a boundary in Columbus, OH, but S_{17} is inside in study region. Also, for log transformed data, we observe that Y_4 and Y_{17} are possibly outliers because those are smaller than 2.5% quantile of the entire data. Thus, we eliminate Y_4 and Y_{17} as outliers. From linear model without considering spatial dependency, we observe that only house value has a significant effect on the log-transformed crime rate. However, when Y_4 and Y_{17} are deleted as outliers, we observe that only house income has a significant effect based on a linear model. From the estimated correlogram in Figure 2, it appears that spatial correlations are not as strong. We model the log-transformed crime rate with income and housing values as the explanatory variables deleting the outliers.

For this data set, we denote $Z_i = \log(Y_i)$ for $i = 1, 2, \ldots, n$, where Y_i is the total of residential burglaries and vehicle thefts per thousand households of sub-area i in 1980. Thus, Z_i is a log transformed crime rate of sub-area S_i. Also, we denote X_1 as the centered housing value in thousand dollars and X_2 as the centered income in thousand dollars, and let $\mathbf{X} = (1, X_1, X_2)$ denote the vector of predictions. With this data set, we consider linear regression model with iid errors and correlated errors (modeled

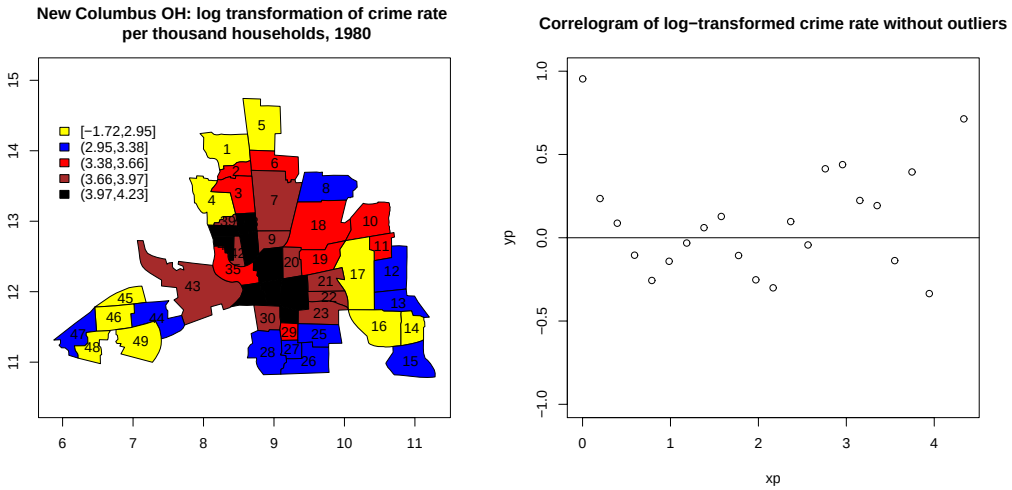

Figure 2: The log-transformed crime rate of 49 neighborhoods in Columbus, OH, 1980 that are divided into 5 intervals of each 20% quantiles and correlogram of log transformed crime rate of Columbus OH deleting outliers

by CAR and DCAR process). We obtain the MLE's of ρ, σ^2, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ and $\boldsymbol{\delta} = (\delta_1, \delta_2)$ under different modeling assumptions. We consider the following models:

$$Z_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i \quad i = 1, \ldots, n$$

Model 1. $\epsilon_i \sim N(0, \sigma^2)$: iid errors

Model 2. $\boldsymbol{\epsilon} \sim N\big(\mathbf{0}, \sigma^2(\mathbf{I} - \rho\tilde{\mathbf{W}})^{-1}\mathbf{D}\big)$: CAR errors

Model 3. $\boldsymbol{\epsilon} \sim N\big(\mathbf{0}, \sigma^2(\mathbf{I} - \delta_1\mathbf{W}^{(1)} - \delta_2\mathbf{W}^{(2)})^{-1}\mathbf{D}\big)$: DCAR errors

Parameter estimates of these models are displayed in Tables 4 and 5. From Table 4, we observe that the house value is not significant at 5% level under Model 1. From Table 5, we observe that $\widehat{\rho}_{MLE} = -0.456$ but standard error (SE) of estimates of ρ is 0.493. It means that there is negative spatial dependence for the log-transformed crime rate, but the spatial dependence is not strong under CAR assumption. However, $\widehat{\delta}_{1MLE} =$

-0.095 and $\widehat{\delta}_{2MLE} = -0.896$ with SE's 0.607 and 0.566, respectively. It means that spatial dependence in the NE/SW direction is not strong, but spatial dependence in the NW/SE direction is strongly negative. This clearly demonstrates the advantage of DCAR model in estimating the directional specific spatial dependence in contrast to isotropic CAR model which would have concluded weak spatial correlation.

Table 4: Linear model of house value and income on log transformed crime rate of Columbus OH without outliers (Model 1)

Coefficients	Estimate	Std.Err.	t-value	p-value
β_0	3.486	0.040	88.098	< 0.001
β_1(house value)	-0.003	0.003	-1.010	0.318
β_2(income)	-0.066	0.009	-7.463	< 0.001
	R^2: : 0.660	Adj R^2: 0.645	Residual standard error: 0.270	

Table 5: Estimated crime rate of Columbus OH with CAR and DCAR model for the latent spatial process deleting outliers (Model 2-3)

	CAR		DCAR	
Parameter	Est.	Std.Err.	Est.	Std.Err.
ρ	-0.456	0.493	-	-
δ_1	-	-	-0.095	0.607
δ_2	-	-	-0.896	0.566
σ^2	0.336	0.071	0.320	0.069
β_0	3.488	0.034	3.489	0.033
β_1	-0.002	0.003	-0.002	0.003
β_2	-0.070	0.009	-0.073	0.009
AIC	28.068		29.095	
BIC	37.319		40.196	

To compare models, we consider AIC , BIC and mean squared predicted error based on Leave-one-out cross-validation method and we denote it by MSPE and is defined as

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \widehat{z}_{-i})^2,$$

where $\widehat{z}_{-i}$ is the predicted value of ith area based leaving out the ith observation. For

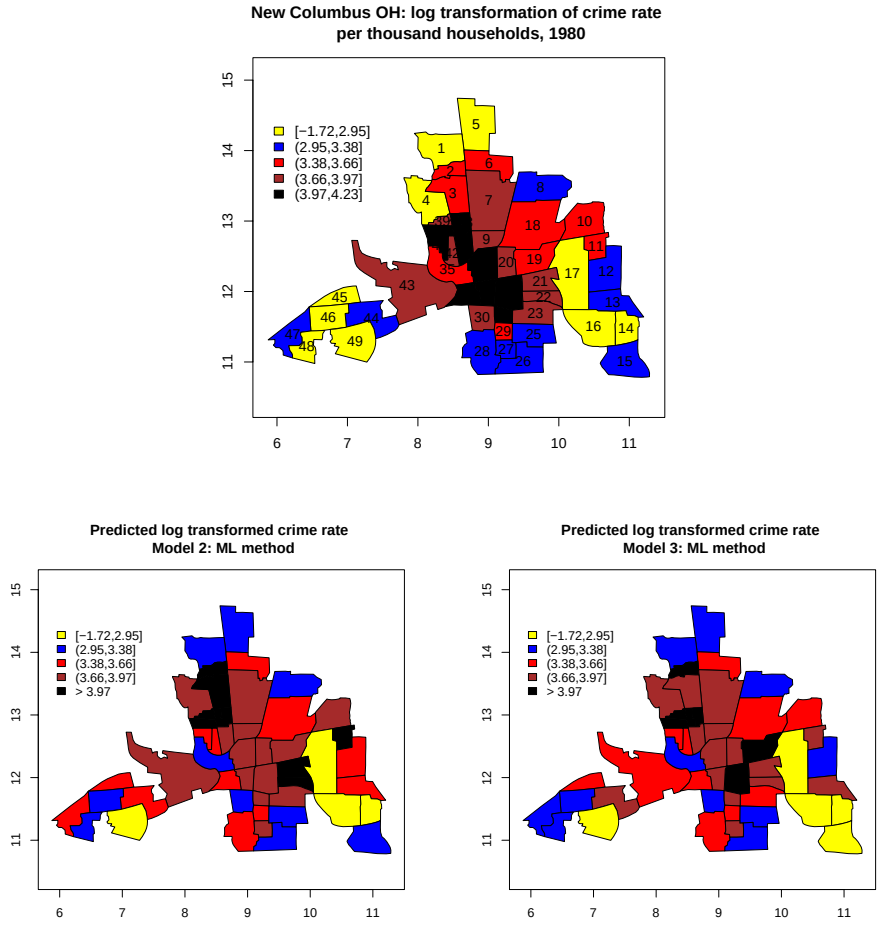

Figure 3: Predicted Log transformed crime rate of Columbus OH (Model 2 and Model 3)

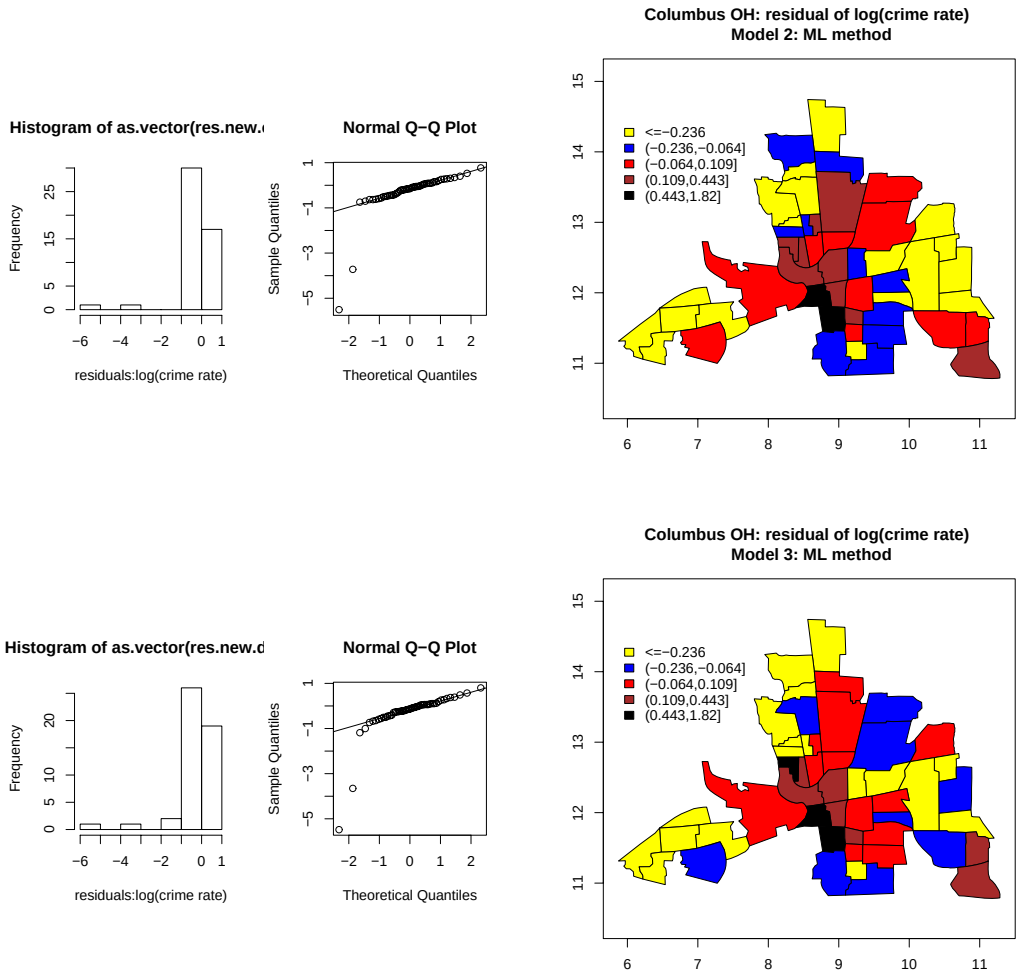

Figure 4: Residual plots of log transformed crime rate of Columbus OH (Model 2 and Model 3)

the predicted values, the best linear unbiased predictor (BLUP) (Schabenberger and Gotaway, 2005) is considered and is given by

$$\text{Model 1. } \mathbf{X}_i\widehat{\boldsymbol{\beta}}$$

$$\text{Model 2. } \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \widehat{\rho}\sum_{j=i}^{n}\tilde{w}_{ij}(z_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}})$$

$$\text{Model 3. } \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \widehat{\delta}_1\sum_{j=1}^{n}\tilde{w}_{ij}^{(1)}\big(z_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}\big) + \widehat{\delta}_2\sum_{j=1}^{n}\tilde{w}_{ij}^{(2)}\big(z_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}\big).$$

First, from Table 5, we observe that AIC and BIC of CAR error model are smaller than those of DCAR. Also, in Table 6, we observe that MSPE of Model 2 is the smallest. However, AIC and BIC of DCAR error model are not quite nominally different from those of CAR error. Also, MSPE of Model 3 (0.037) is very close to that of Model 2 (0.036). From Figure 3 and 4, we observe that the predicted value based on CAR error model appears to smooth out the log transformed crime rate over Columbus, OH. Also, from residual plots of Model 2 (CAR error model), we observe that residuals form blocks in the NW/SE direction. However, the predicted value based on CAR error model appears to have higher rate in NW/SE direction than NE/SW direction. From residual plots of Model 3 (DCAR error model), we observe that residuals seem not to have any trend over study area. Therefore, we conclude that the log-transformed crime rate of Columbus, OH has different spatial dependence for neighbors in NW/SE direction and that of NE/SW direction. This matches with the conclusion of Anselin who used two separate models to capture the directional effects.

	MSPE
Model 1	0.080
Model 2	0.036
Model 3	0.037

Table 6: Mean Squared Predicted Error of Leave-one-out method (MSPE_{L})

5 Extensions and future work

DCAR model as an extension of CAR model, captures the directional spatial dependence in addition to distance specific correlation. The DCAR model is also found to be as efficient as the CAR model when data are generated from the CAR model. However, CAR models usually fail to capture the directional effects when data is generated from DCAR or other anisotropic model. Our model proposed in (6) can be extended to more than two directions and one may similarly fit models of the type

$$\boldsymbol{Z} \sim N_n\big(\mathbf{X}\boldsymbol{\beta}, \sigma^2(\mathbf{I} - \sum_{k=1}^{M} \delta_k \tilde{\mathbf{W}}^{(k)})^{-1}\mathbf{D}\big),$$

where $\tilde{\mathbf{W}}^{(k)}$ denote the matrices of weights specific to k-th directional effect. Our current work involves the creation of even generated model that estimate the $\tilde{\mathbf{W}}^{(k)}$'s adaptively from the data and will be published else where.

Appendix A: Lemmas

Lemma 2 *Let* $\mathbf{A} = \big(a_{ij}\big)$ *be a* $n \times n$ *matrix. If* $a_{ii} > \sum_{j \neq i} |a_{ij}|$ *for all* i*, then* $\mathbf{A}$ *is positive definite.*

Proof: See Ortega, 1987, P.226. □

Lemma 3 *Jacobi's formula: Let* $\mathbf{B}$ *be* $n \times n$ *matrix. Then,* $d|\mathbf{B}| = \mathrm{Tr}(\mathrm{Adj}(\mathbf{B})d\mathbf{B})$*, in which* $\mathrm{Adj}(\mathbf{B})$ *is the adjugate of the matrix* $\mathbf{B}_{n\times n}$ *and* $d\mathbf{B}$ *is its differentiable, where* $|\cdot|$ *is determinant of matrix.*

Proof: Given an $n \times n$ matrix $\mathbf{B} = \{b_{ij}\}$, its adjugate $\mathrm{Adj}(\mathbf{B}) = \mathbf{A} = \{a_{ij}\}$ is defined as

$$a_{ij} = (-1)^{i+j}|\mathbf{B} \text{ without its } j\text{th row and } i\text{th column}|$$

so that $\mathbf{AB} = \mathbf{BA} = |\mathbf{B}|\mathbf{I}$. a_{ij} is the cofactor of b_{ji} in $|\mathbf{B}|$, so a_{ij} is a polynomial function of the elements of $\mathbf{B}$ but independent of b_{jk} and b_{ki} for all k. In $|\mathbf{B}| = \sum_k b_{ik}a_{ki}$, each element b_{ij} is linearly multiplied by its cofactor a_{ji}, so $\partial|\mathbf{B}|/\partial b_{ij} = a_{ji}$. Thus,

$$\mathrm{d}|\mathbf{B}| = \sum_j \sum_i (\partial|\mathbf{B}|/\partial b_{ij})\mathrm{d}b_{ij} = \sum_j \sum_i a_{ji}\mathrm{d}b_{ij} = \mathrm{Tr}(\mathrm{Adj}(\mathbf{B}) \cdot \mathrm{d}\mathbf{B}).\square$$

Lemma 4 *If* $\mathrm{Rank}(\mathbf{B}_{n\times n}) = n$ *(full rank),* $\mathrm{d}\log|\mathbf{B}| = \mathrm{Tr}(\mathbf{B}^{-1}\mathrm{d}\mathbf{B})$.

Proof: See Section 10.8 of Graybill (1969). $\square$

Lemma 5 *Let* $\mathbf{A}$ *be an* $n \times n$ *matrix with all real characteristic roots and let exactly* h $(0 < h \leq n)$ *of them be nonzero; then*

$$[tr(\mathbf{A})]^2 \leq h \cdot tr(\mathbf{A}^2).$$

Proof of Lemma: See p.228-229 of Graybill (1969). $\square$

Appendix B: Proof of Theorem 2

To show that MLE of Gaussian DCAR model is asymptotically normal and weakly consistent, we assume that $\mathbf{X}$ is suitably chosen so that $\lim(\mathbf{X}^T\mathbf{X})^{-1} = 0$ to hold and

hence Condition (iv) of Theorem 1 holds. We note that $\tilde{\mathbf{W}}$ can be expressed as $\mathbf{DW}$. Similarly, $\tilde{\mathbf{W}}^{(1)} = \mathbf{DW}_1$ and $\tilde{\mathbf{W}}^{(2)} = \mathbf{DW}_2$. We now show that Conditions (i)-(iii) of Theorem 1 are satisfied by the DCAR model.

Condition (i):For DCAR model, the variance covariance matrix $\mathbf{\Sigma}$ is symmetric and positive definite if $\max(|\delta_1|, |\delta_2|) < 1$. Let $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n$ be the eigenvalues of $\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}$. We can express the determinant of $\mathbf{\Sigma}$ as $|\mathbf{\Sigma}| = \sigma^{2n}\prod_{i=1}^n \lambda_i$ and that of $\mathbf{\Sigma}^{-1}$ as $|\mathbf{\Sigma}^{-1}| = \sigma^{-2n}\prod_{i=1}^n \frac{1}{\lambda_i}$. As $\frac{1}{\lambda_1}$ is the largest eigenvalue of $\mathbf{D}^{-1}\mathbf{A}^*(\boldsymbol{\delta})$, we have $\big(\mathbf{D}^{-1} - \delta_1\mathbf{W}_1 - \delta_2\mathbf{W}_2\big)\mathbf{x} = \frac{1}{\lambda_1}\mathbf{x}$, for some $||\mathbf{x}|| = 1$. Thus, $\mathbf{x}^T\mathbf{D}^{-1}\mathbf{x} - \delta_1\mathbf{x}^T\mathbf{W}_1\mathbf{x} - \delta_2\mathbf{x}^T\mathbf{W}_2\mathbf{x} = \frac{1}{\lambda_1}$. As $\mathbf{x}^T\mathbf{D}^{-1}\mathbf{x} = \sum_{i=1}^n m_i x_i^2 \leq m_{(n)}$, so it follows that $\limsup \lambda_n < \infty$. $|\mathbf{\Sigma}_{(1)}| = |\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}| = \prod_{i=1}^n \lambda_i$. Thus as it is shown above, $\lim|\lambda_n^1| = \lim \lambda_n = C = C_1 < \infty$. For $\mathbf{\Sigma}_{(2)}$, $\mathbf{\Sigma}_{(2)} = \sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\tilde{\mathbf{W}}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} = \sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}$. Thus, the determinant of $\mathbf{\Sigma}_{(2)}$ is $|\mathbf{\Sigma}_{(2)}| = \sigma^{2n}|\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}||\mathbf{W}^{(1)}||\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}|$
$= \sigma^{2n}\big(\prod_{i=1}^n \lambda_i^2|\xi_i^{(1)}|\big)$, where $\xi_i^{(1)}$'s are eigenvalues of symmetric matrix $\mathbf{W}^{(1)}$. Here, based on Theorem 1 in Lancaster and Tismenetsky (1985) p.359 that is for any matrix norm, if $\mathbf{A} \in \mathbb{R}^{n\times n}$ and λ_A is the eigenvalue of $\mathbf{A}$ then $||\mathbf{A}|| \geq \lambda_A$, it can be shown that $|\xi_n^{(1)}| \leq ||W_1|| = \sum_{i,j=1}^n w_{1ij} \leq \max(m_i)$, so $|\xi_n^{(1)}| \leq M^*$, where $|\xi_n^{(1)}| = \max(|\xi_i^{(1)}|)$. Thus, $\lim|\lambda_n^2| = \lim \sigma^{2n}\lambda_n^2|\xi_n^{(1)}| = C_2 < \infty$. Similarly, for $\mathbf{\Sigma}_{(3)}$, $\mathbf{\Sigma}_{(3)} = \sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\tilde{\mathbf{W}}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} = \sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}$. Thus, similar with $\mathbf{\Sigma}_{(2)}$, $\lim|\lambda_n^3| = \lim \sigma^2\lambda_n^2|\xi_n^{(2)}| = C_3 < \infty$, where $\xi_i^{(2)}$'s are eigenvalues of $\mathbf{W}^{(2)}$. Also, for

the eigenvalues of $\boldsymbol{\Sigma}_{(ij)}$, we derive that

$$\begin{aligned}
\boldsymbol{\Sigma}_{(11)} &= \mathbf{0} \\
\boldsymbol{\Sigma}_{(12)} &= \boldsymbol{\Sigma}_{(21)} = \mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} \\
\boldsymbol{\Sigma}_{(13)} &= \boldsymbol{\Sigma}_{(31)} = \mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} \\
\boldsymbol{\Sigma}_{(22)} &= 2\sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} \\
\boldsymbol{\Sigma}_{(33)} &= 2\sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D} \\
\boldsymbol{\Sigma}_{(23)} &= 2\sigma^2\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}.
\end{aligned}$$

Thus, based on above results,

$$\begin{aligned}
& \lim|\lambda_n^{11}| = 0 \\
|\boldsymbol{\Sigma}_{(12)}| &= |\boldsymbol{\Sigma}_{(21)}| = \prod_{i=1}^{n}\lambda_i^2|\xi_i^{(1)}| \Rightarrow \lim|\lambda_n^{12}| = \lim|\lambda_n^{21}| = C_{12} = C_{21} = C_2 < \infty \\
|\boldsymbol{\Sigma}_{(13)}| &= |\boldsymbol{\Sigma}_{(31)}| = \prod_{i=1}^{n}\lambda_i^2|\xi_i^{(2)}| \ \Rightarrow \lim|\lambda_n^{13}| = \lim|\lambda_n^{31}| = C_{13} = C_{31} = C_3 < \infty \\
|\boldsymbol{\Sigma}_{(22)}| &= (2\sigma^2)^n\prod_{i=1}^{n}\lambda_i^3|\xi_i^{(1)}|^2 \Rightarrow \lim|\lambda_n^{22}| = \lim 2\sigma^2\lambda_n^3|\xi_n^{(1)}|^2 = C_{22} < \infty \\
|\boldsymbol{\Sigma}_{(33)}| &= (2\sigma^2)^n\prod_{i=1}^{n}\lambda_i^3|\xi_i^{(2)}|^2 \Rightarrow \lim|\lambda_n^{33}| = \lim 2\sigma^2\lambda_n^3|\xi_n^{(2)}|^2 = C_{33} < \infty \\
|\boldsymbol{\Sigma}_{(23)}| &= |\boldsymbol{\Sigma}_{(32)}| = (2\sigma^2)^n\prod_{i=1}^{n}\lambda_i^3|\xi_i^{(1)}||\xi_i^{(2)}| \\
&\Rightarrow \lim|\lambda_n^{22}| = \lim 2\sigma^2\lambda_n^3|\xi_n^{(1)}||\xi_n^{(2)}| = C_{23} = C_{32} < \infty.
\end{aligned}$$

Condition (ii):For the Euclidean norm of $\boldsymbol{\Sigma}_{(i)}$, $||\boldsymbol{\Sigma}_{(i)}||^2 \leq \sum_{l=1}^{n}(\lambda_l^i)^2 \leq n(\lambda_n^i)^2$. Thus, $||\boldsymbol{\Sigma}_{(i)}||^{-2} \geq \frac{1}{n(\lambda_n^i)^2}$. Here, λ_n^i converges to a finite constant as $n \to \infty$ for all $i = 1, \ldots, k$ for $\alpha = \frac{1}{2}$. Therefore, $||\boldsymbol{\Sigma}_{(i)}||^{-2} = O(n^{-\frac{1}{2}-\alpha})$ for all $i = 1, \ldots, k$.

Condition (iii): For the third condition of Theorem 1, $t_{ij} = tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(i)}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{(j)})$ for

$i, j = 1, \ldots, k$ are

$$
\begin{aligned}
t_{11} &= n\sigma^{-4} \\
t_{12} &= t_{21} = \sigma^{-2} tr(\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}) \quad t_{13} = t_{31} = \sigma^{-2} tr(\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}) \\
t_{22} &= tr(\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}) \quad t_{33} = tr(\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}) \\
t_{23} &= t_{32} = tr(\mathbf{W}^{(1)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}\mathbf{W}^{(2)}\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}).
\end{aligned}
$$

For the ease of notation, let $\mathbf{T}_l = \mathbf{W}_l\mathbf{A}^*(\boldsymbol{\delta})^{-1}\mathbf{D}$ for $l = 1, 2$. Then,

$a_{ii} = \lim\{t_{ii}/(t_{ii}t_{ii})^{1/2}\} = 1$ for $i = 1, 2, 3$

$a_{12} = a_{21} = \lim\left\{\frac{tr(\mathbf{T}_1)^2}{ntr(\mathbf{T}_1\mathbf{T}_1)}\right\}^{1/2} \Rightarrow a_{12} = a_{21} = \lim\left\{\frac{1}{\frac{ntr(\mathbf{T}_1\mathbf{T}_1)}{tr(\mathbf{T}_1)^2}}\right\}^{1/2}$.
Here, by Lemma 5, $\frac{tr(\mathbf{T}_1\mathbf{T}_1)}{tr(\mathbf{T}_1)^2} \geq \frac{1}{h}$ $(0 < h \leq n)$. Thus, $a_{12} = a_{21} = 0$.

$a_{13} = a_{31} = \lim\left\{\frac{1}{\frac{ntr(\mathbf{T}_2\mathbf{T}_2)}{tr(\mathbf{T}_2)^2}}\right\}^{1/2}$.
Similarly, by Lemma 5, $a_{13} = a_{31} = 0$.

$a_{23} = a_{32} = \lim\left\{\frac{1}{\frac{tr(\mathbf{T}_1\mathbf{T}_1)tr(\mathbf{T}_2\mathbf{T}_2)}{tr(\mathbf{T}_1\mathbf{T}_2)^2}}\right\}^{1/2}$. For denominator, we consider eigenvalues of $\mathbf{A}^*(\boldsymbol{\delta})\mathbf{D}$ and $\mathbf{W}_l$ for $l = 1, 2$, then we express the denominator as

$$
\begin{aligned}
\left\{\frac{tr(\mathbf{T}_1\mathbf{T}_1)tr(\mathbf{T}_2\mathbf{T}_2)}{tr(\mathbf{T}_1\mathbf{T}_2)^2}\right\}^{1/2} &= \frac{\sum_{i=1}^n \lambda_i^2(\xi_i^{(1)})^2 \sum_{i=1}^n \lambda_i^2(\xi_i^{(1)})^2}{\sum_{i=1}^n \lambda_i^2|\xi_i^{(1)}||\xi_i^{(2)}|} \\
&\leq \frac{n\lambda_n^2|\xi_n^{(1)}| \cdot n\lambda_n^2|\xi_n^{(2)}|}{\lambda_1^2\sum_{i=1}^n |\xi_i^{(1)}||\xi_i^{(2)}|} = n^2\Big(\frac{\lambda_n}{\lambda_1}\Big)^2 \frac{1}{\sum_{i=1} \frac{|\xi_i^{(1)}|}{|\xi_n^{(1)}|}\frac{|\xi_i^{(2)}|}{|\xi_n^{(2)}|}},
\end{aligned}
$$

where, $|\xi_n^{(l)}| = \max(|\xi_i^{(l)}|)$ for $l = 1, 2$. As $n \to \infty$, $n^2\Big(\frac{\lambda_n}{\lambda_1}\Big)^2 \frac{1}{\sum_{i=1} \frac{|\xi_i^{(1)}|}{|\xi_n^{(1)}|}\frac{|\xi_i^{(2)}|}{|\xi_n^{(2)}|}} \to \infty$,
because $\frac{|\xi_i^{(l)}|}{|\xi_n^{(l)}|} \leq 1$ for all $l = 1, 2$. Thus, $a_{23} = a_{32} = 0$.

Therefore, $\mathbf{A} = [a_{ij}] = \mathbf{I}_{3\times 3}$ is an identity matrix, so it is nonsingular. Thus by Theorem 2, we note that MLE's of Gaussian DCAR model are consistent and asymptotically normal. □

References

Akaike, H. (1974) "A new look at statistical model identification," *IEEE Transactions on Automatic Control* 19, 716-723.

Anselin, Luc. (1988) *Spatial econometrics: methods and models*, Kluwer Academic Publishers.

Besag, J. (1974) "Spatial interaction and the statistical analysis of lattice systems" (with discussion), *Journal of the Royal Statistical Society, Series B* 36, 192-236.

Breslow, N. E. and Clayton, D. G. (1993) "Approximate inference in generalized linear mixed models," it Journal of the American Statistical Association 88, 9-25.

Brook, D. (1964) "On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems," *Biometrika* 51, 481-483.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) "A limited memory algorithm for bound constrained optimization," *SIAM Journal of Scientific Computing* 16,1190-1208.

Cliff, A. D. and Ord, J. K. (1981) *Spatial Processes: Models & Applications*, Pion Limited.

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

Efron, B. and Hinkley, D. V. (1978) "Assessing the accuracy of the maximum likelihood estimator: observed versus expected fisher information," *Biometrika* 65, 457-

482.

Graybill, F. A. (1969) *Introduction to Matrices with Applications in Statistics*, Wadsworth Publishing Company.

Lancster, P. and Tismenetsky, M. (1985) *The Theory of Matrices*, Second Edition, Academic Press.

Mardia, K. V. and Marshall, R. J. (1984) "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika* 71, 135-146.

Miller, H. J. (2004) "Tobler's first law and spatial analysis," *Annals of the Association of American Geographers* 94, 284-295.

Ord, K. (1975) "Estimation methods for models of spatial interaction," *Journal of the American Statistical Association* 70, 120-126.

Ortega, J. M. (1987) *Matrix Theory*, New York:Plenum Press.

Schabenberger, O. and Gotaway, C. A. (2005) *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC.

Schwartz, G. (1978) "Estimating the dimension of a model," *Annals of Statistics* 6, 461-464.

Stone, M. (1977) "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion," *Journal of the Royal Statistical Society, Series B* 39, 44-47.

Sweeting, T., J. (1980) "Uniform asymptotic normality of the maximum likelihood estimator," *Annals of Statistics* 8, 1375-1381.

van der Linde, A., Witzko, K.-H. and Jockel, K.-H. (1995) "Spatio-temporal analysis of mortality using splines" *Biometrics* 4, 1352-1360.

Wahba, G. (1977) "Practical approximate solutions to linear operator equations when the data are noisy," *SIAM Journal on Numerical Analysis* 14, 651-667.