Performance of Information Criteria for Spatial Models

Hyeyoung Lee* and Sujit K. Ghosh*

Institute of Statistics Mimeo Series #2588

Abstract

Model choice is one of the most crucial aspect in any statistical data analysis. It is well known that most models are just an approximation to the true data generating process but among such model approximations it is our goal to select the "best" one. Researchers typically consider a finite number of plausible models in statistical applications and the related statistical inference depends on the chosen model. Hence model comparison is required to identify the "best" model among several such candidate models. This article considers the problem of model selection for spatial data. The issue of model selection for spatial models has been addressed in the literature by the use of traditional information criteria based methods, even though such criteria have been developed based on the assumption of independent observations. We evaluate the performance of some of the popular model selection critera via Monte Carlo simulation experiments using small to moderate samples. In particular, we compare the performance of some of the most popular information criteria such as Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Corrected AIC (AICc) in selecting the true model. The ability of these criteria to select the correct model is evaluated under several scenarios. This comparison is made using various spatial covariance models ranging from stationary isotropic to nonstationary models.
Keywords: model selection, spatial models, information criteria

*Department of Statistics, North Carolina State University, Raleigh, NC 27695, U.S.A.

1 Introduction

Model selection is an important part of any statistical analysis. In fact, G.E.P. Box once remarked that "*Models, of course, are never true, but fortunately it is only necessary that they be useful*" (Box, 1979, p.2). In this sense among all the approximating models, we would like to use the one that is "closest" to the true data generating process. In practice, most researchers often consider a given set of plausible models postulated and formulated based on the background knowledge and preliminary data analysis. As the choice of one model over the another can make a substantial difference in statistical inference, we need a "good" method to select among a set of postulated models. Thus a model comparison tool is required to identify the "best" model (if possible) among several candidate models. Many researchers have examined this issue and various methods for selecting the "best " model have been suggested. Examples include hypothesis testing, cross validation (Stone, 1974), R^2, Mallows' C_p (Mallows, 1973), and several information criteria.

Information criteria are probably one of the most widely used tools in model selection. However, there have been few investigations of the performance of these criteria in a spatial modeling context. Hoeting et al. (2006) discussed the issue of model selection for geostatistical data. They explored the effect of spatial correlation on variable selection using Akaike Information Criterion (AIC) (Akaike, 1973) in geostatistical models. Their simulation results showed that spatially Corrected AIC (AICc) (Sugiura, 1978; Hurvich and Tsai, 1989) outperforms independent AICc which ignores spatial correlation in the variable selection.

In particular, few studies have compared the performance between different information criteria like AIC and Bayesian Information Criterion (BIC) (Schwarz, 1978). Hence, little is known about the relative performance of different information criteria. Currently, no consensus exists on the best criterion for spatial model selection. Of particular interest is how these different criteria perform with various spatial covariance models. We explore this issue via extensive Monte Carlo simulation experiments using a wide variety of spatial models. The purpose of this study is to examine the performance of different information criteria for use in spatial covariance model selection. We compare the performance of traditional information criteria such as AIC, BIC, and AICc. This comparison is made using various spatial covariance models ranging from stationary isotropic to nonstationary models.

The remainder of this article is organized as follows. Section 2 provides brief descriptions of various information criteria used for model selection such as AIC, BIC and AICc. In Section 3, we describe various spatial covariance models such as stationary isotropic and anisotropic, and nonstationary models that will be used as an illustration to generate data from a specific covariance model. Section 4 presents the results from simulations which compare the performance of AIC, BIC and AICc with regard to their ability to identify the true model among various spatial covariance models. Discussions and future research are summarized in Section 5.

2 Information Criteria for Model Selection

Information criteria have played an important role in model selection. These are based on the Kullback-Leibler (K-L) information (Kullback and Leibler, 1951) which is defined as

$$I(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x|\theta)}\right)dx,$$

where θ represents parameters in model g. Here, $I(f,g)$ can be interpreted as the "information lost when the model g is used to approximate full reality or truth f" (Burnham and Anderson, Section 2.1, 2002). Hence, the model that minimizes $I(f,g)$ will be considered as the best model. However, $I(f,g)$ cannot be used directly in model selection because $f(x)$ and θ are not known.

A variety of information criteria have been proposed to be used in model selection. These include AIC, AICc, Takeuchi's Information Criterion (TIC) (Takeuchi, 1976), and QAIC and QAICc (Quasi-likelihood modifications to AIC and AICc) (Lebreton et al., 1992). These criteria are estimates of the relative K-L information between f and g and are based on the concept that a true f may not be included in the set of models being evaluated (Burnham and Anderson, 2002). On the other hand, several criteria such as Bayesian Information Criterion (BIC) (Schwarz, 1978) have been developed assuming that a true f belongs to the set of models being considered as candidate models (Burnham and Anderson, 2002). Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) is analogous to AIC from a Bayesian perspective.

A substantial advantage in using information criteria is that these are valid even when the models being compared and not nested. Traditional

likelihood ratio tests are defined only for nested models, and this represents a limitation in the use of hypothesis testing in model selection.

Most information criteria have a form that consists of two terms. In general, the first term is the negative log-likelihood, multiplied by two, of the data calculated with the maximum likelihood estimates of the parameters. The second term differs between different information criteria. The second term is often interpreted as a penalty for model complexity. Hence, it increases as the number of parameters in the model increases.

In model selection using information criteria, the model that minimizes information criteria is declared as the best model among the set of models under consideration. In the following sections, we describe commonly used information criteria such as AIC, AICc, and BIC.

2.1 Akaike Information Criterion (AIC)

Akaike Information Criterion (AIC) (Akaike, 1973) is one of the most well known information criteria used in model selection. AIC is an estimate of relative, expected K-L information (Kullback and Liebler, 1951) between a fitted model and the true model. AIC is defined as

$$\text{AIC} = -2\log\left[L(\hat{\theta}|X)\right] + 2p, \tag{1}$$

where $\log L(\hat{\theta}|X)$ represents the log-likelihood function of the maximum likelihood estimator (MLE), $\hat{\theta}$, given the observed data X, and p is the dimension of the parameter θ. The first term can be interpreted as a measure of lack of model fit, while the second term can be interpreted as a penalty for increasing the dimension of the model. The second term is the asymptotic bias-correction term derived from an asymptotic estimator of relative, expected K-L information (Burham and Anderson, 2002).

In application, we compute AIC for each of the candidate models and selects the model with the smallest value of AIC. Models producing smaller values of AIC can be thought of as having a smaller difference from the true model. AIC provides a simple and effective means for the selection of the best approximating model to the true model (Burnham and Anderson, 2002).

With regard to general linear models, AIC is known to perform relatively well for small samples, however the criterion does not tend to select the true model in large samples (Hurvich and Tsai, 1990).

2.2 Corrected Akaike Information Criteria (AICc)

As an approximately unbiased estimator of the expected K-L information of a fitted model, AIC has been shown to be strongly negatively biased in small samples (Sugiura, 1978; Hurvich and Tsai, 1989). Hurvich and Tsai (1989) derived a bias-corrected version of AIC, AICc. They argued that AICc should be used in place of AIC, when the dimension of the model is large relative to sample size or when n is small, for any p. The AICc is defined as

$$\begin{aligned} \text{AICc} &= -2\log\left[L(\hat{\theta}|X)\right] + 2p\left(\frac{n}{n-p-1}\right) \\ &= -2\log\left[L(\hat{\theta}|X)\right] + 2p + \frac{2p(p+1)}{n-p-1} \\ &= \text{AIC} + \frac{2p(p+1)}{n-p-1}, \end{aligned} \tag{2}$$

where n is the sample size and p is the number of parameters in the model. AICc has an additional bias-correction term compared to AIC, which is adjusted to the parameter complexity p and the sample size n. However, if n is large with respect to p, then this additional bias-correction is negligible and AIC should perform well. Burnham and Anderson (2002) advocated the use of AICc, in particular, when the ratio $n/p < 40$ for the model with the largest value of p. If n/p is sufficiently large, then AIC and AICc are similar and will tend to select the same model. They also mentioned that AICc should be used in practice, because AICc converges to AIC as n gets large, with p fixed.

2.3 Bayesian Information Criterion (BIC)

Along with AIC, Bayesian Information Criterion (BIC) (Schwarz, 1978) is currently among the most commonly used information criteria in model selection. BIC is usually explained in terms of Bayesian theory, especially as an approximation of the Bayes factor, which is the ratio of the marginal likelihoods for two models. Unlike AIC, BIC is not an estimate of relative expected K-L information (Burnham and Anderson, 2002). BIC is defined as

$$\text{BIC} = -2\log\left[L(\hat{\theta}|X)\right] + p\log(n), \tag{3}$$

where $\log\left[L(\hat{\theta}|X)\right]$ again represents the log-likelihood function of $\hat{\theta}$, which is the maximum likelihood estimator (MLE) based on the observed data X; p is the number of parameters in the model, and n is the sample size. The first term of BIC is same as that of AIC. However, the second term penalizes the model with increased model complexity, or larger p, and sample size as well. AIC and BIC differ only by the coefficient multiplying the number of parameters, in other words, by how strongly they penalize large models. In general, models chosen by BIC are more parsimonious than those chosen by AIC. As usually used, one computes the BIC for each model and selects the model with the smallest criterion value. In contrast to AIC, BIC tends to choose the true model in large samples. However, BIC has also known to perform poorly in small samples in the context of general linear models (Hurvich and Tsai, 1990).

3 Spatial Models

We consider various geostatistical models that are popularly used for point-referenced data. In particular, we evaluate the performance of information criteria using models that range from stationary (including anisotropic) to nonstationary models.

3.1 Stationary Processes

Consider a random process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, where D is a fixed subset of $\Re^d$. Assume that the random process $Z(\cdot)$ satisfies

$$E(Z(\boldsymbol{s})) = \mu, \qquad \text{for all } \boldsymbol{s} \in D \quad \text{and} \tag{4}$$

$$Cov(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = C(\boldsymbol{s}_i - \boldsymbol{s}_j), \qquad \text{for all } \boldsymbol{s}_i, \boldsymbol{s}_j \in D. \tag{5}$$

That is, the mean does not depend on $\boldsymbol{s}$ and the covariance is a function only of the increment $\boldsymbol{s}_i - \boldsymbol{s}_j$. Then $Z(\cdot)$ is said to be a *second-order* or *weak stationary* process. Furthermore, if $C(\boldsymbol{s}_i - \boldsymbol{s}_j)$ is a function of $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ only, that is, the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, then $C(\cdot)$ is called *isotropic*. An isotropic process assumes that the correlation structure between sites is circular which indicates the correlation depends only on the distance between sites.

One frequently used isotropic covariance function is the exponential model. Here the covariance between measurements at two locations is an exponential function of the distance between two locations,

$$Cov(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = \sigma^2 \exp(-\phi \|\boldsymbol{s}_i - \boldsymbol{s}_j\|) + \tau^2 I(i = j),\ \sigma^2 > 0, \phi > 0, \tau^2 > 0, \quad (6)$$

where $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is the distance between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, and I denotes the indicator function. Here σ^2 and ϕ are positive parameters called the partial sill and the decay or inverse range parameter, respectively. When $i = j$, $d_{ij} = 0$ and $C(d_{ij}) = Var(Z(s_i))$ is often expanded to $\tau^2 + \sigma^2$, where $\tau^2 > 0$ is called a nugget effect, and $\tau^2 + \sigma^2$ is called the sill.

Many other parametric models for the isotropic covariance function are also commonly used (Schabenberger and Gotway, 2005, Section 2.1). Isotropic processes are popular because a number of relatively simple parametric forms are available.

If dependence between $Z(\boldsymbol{s}_i)$ and $Z(\boldsymbol{s}_j)$ is a function of both the distance and the direction of $\boldsymbol{s}_i - \boldsymbol{s}_j$, then the process $\boldsymbol{Z}$ is called anisotropic. Hence, the covariance function, $C(\boldsymbol{s}_i - \boldsymbol{s}_j)$ is no longer purely a function of distance between two spatial locations, $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

Sometimes the anisotropy can be corrected by a linear transformation of the increment vector $\boldsymbol{s}_i - \boldsymbol{s}_j$. This anisotropy is known as *geometric anisotropy* and gives elliptical contours for the correlation. Specifically, the geometric anisotropy is corrected by (i) a rotation of the coordinate system to align the major and minor axes of the elliptical contours, and (ii) a compression of the major axis to make the contours spherical. Following Schabenberger and Gotway (2005, p.151), the anisotropy matrix A is thus defined as,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} \cos\gamma & -\sin\gamma \\ \sin\gamma & \cos\gamma \end{bmatrix}, \quad (7)$$

where λ and θ are the anisotropy ratio for compression and the anisotropy angle for rotation, respectively. Here λ equals the ratio of the ranges in the directions of the major and minor axes of the elliptical contours. Geometric anisotropy is common for processes that evolve along particular directions. For example, airborne pollution emitted from an industrial plant will likely evolve along the wind directions (Schabenberger and Gotway, 2005, p.151).

In general, geometric anisotropy can be incorporated in the isotropic model by correcting distances. For instance, we can incorporate geometic

anisotropy in the exponential model (6),

$$Cov(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = \sigma^2 \exp(-\phi \|A(\boldsymbol{s}_i - \boldsymbol{s}_j)\|) + \tau^2 I(i = j), \tag{8}$$

where A is the anisotropy matrix in (7).

3.2 Nonstationary Processes

We consider a class of parametric nonstationary covariance models proposed by Hughes-Oliver et al. (1998). They incorporate nonstationarity in the covariance model driven by a point source, (e.g., the center of a wafer in semiconductor processing). Their covariance model for a point source at location $\boldsymbol{c}$ is

$$Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h_{ij} \exp\{\alpha |c_i - c_j| + \beta \min[c_i, c_j]\}\} + \tau^2 I(i = j), \tag{9}$$

where $h_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|, c_i = \|\boldsymbol{s}_i - \boldsymbol{c}\|$, and $c_j = \|\boldsymbol{s}_j - \boldsymbol{c}\|$. Here c_i and c_j are the distances of sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ from the point source $\boldsymbol{c}$, respectively, and α, $\beta \geq 0$. This covariance model is nonstationary because the correlation between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ depends on the distances between sites and the point source through c_i and c_j.

The covariance model (9) can be thought of as a generalization of the exponential model for an isotropic process. Note that when $\alpha = \beta = 0$ in (9), the covariance model (9) reduces to the exponential model (6). Here (9) assumes that the effects of the point source are circular, that is, point source isotropy. Hence the correlation depends only on the distance between sites and on the distance between a site and the point source.

We can also incorporate point source anisotropy in the nonstationary point source isotropic model in a similar way as shown in (8). Schabenberger and Gotway (2005, p.423) presented point source anisotropy incorporated in the model (9),

$$Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h^*_{ij} \exp\{\alpha |c^*_i - c^*_j| + \beta \min[c^*_i, c^*_j]\}\} + \tau^2 I(i = j), \tag{10}$$

where $h^*_{ij} = \|A(\boldsymbol{s}_i - \boldsymbol{s}_j)\|$, $c^*_i = \|A_c(\boldsymbol{s}_i - \boldsymbol{c})\|$, $c^*_j = \|A_c(\boldsymbol{s}_j - \boldsymbol{c})\|$ and A, A_c are the anisotropy matrices in (7).

4 A Simulation Study

In this simulation study, we evaluate and compare the performance of information criteria presented in Section 2 in selecting the models presented

in Section 3. Of particular interest is how these criteria perform with different spatial covariance models. Specifically, we compare the performance of these information criteria with regard to their ability to discriminate the true model under various spatial covariance models, parameter values, and sample sizes.

4.1 Covariance Models

We consider following four different forms of exponential models for spatial covariance functions.

i) Σ_1: Exponential Isotropic Model,

$$\Sigma_{ij} = Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h_{ij}\} + \tau^2 I(i=j),$$

ii) Σ_2: Exponential Anisotropic Model,

$$\Sigma_{ij} = Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h^*_{ij}\} + \tau^2 I(i=j),$$

iii) Σ_3: Exponential Point Source Isotropic Model,

$$\Sigma_{ij} = Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h_{ij} \exp\{\alpha |c_i - c_j|\}\} + \tau^2 I(i=j),$$

iv) Σ_4: Exponential Point Source Anisotropic Model,

$$\Sigma_{ij} = Cov[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)] = \sigma^2 \exp\{-\phi h^*_{ij} \exp\{\alpha |c^*_i - c^*_j|\}\} + \tau^2 I(i=j),$$

where $h_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, $h^*_{ij} = \|A(\boldsymbol{s}_i - \boldsymbol{s}_j)\|$, $c_i = \|\boldsymbol{s}_i - \boldsymbol{c}\|$, $c^*_i = \|A_c(\boldsymbol{s}_i - \boldsymbol{c})\|$, $c_j = \|\boldsymbol{s}_j - \boldsymbol{c}\|$, $c^*_j = \|A_c(\boldsymbol{s}_j - \boldsymbol{c})\|$, and A, A_c are anisotropy matrices in (7).

We use a Gaussian process with the above covariance models to enable likelihood inference. The most convenient assumption would be a multivariate normal distribution for the observed data. That is, suppose we have observations $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \cdots, Z(\boldsymbol{s}_n))'$ at known locations $\boldsymbol{s}_i$, $i = 1, \cdots, n$. We then assume that

$$\boldsymbol{Z}|\boldsymbol{\theta} \sim N_n(\boldsymbol{0}, \Sigma(\boldsymbol{\theta})), \tag{11}$$

where N_n denotes the n-dimensional normal distribution, with mean $\boldsymbol{0}$ and covariance $(\Sigma(\boldsymbol{\theta}))$, where $(\Sigma(\boldsymbol{\theta}))$ takes one of the four forms described above.

Now we consider the following four different spatial models for our simulation studies.

Model	p	$\boldsymbol{\theta}$
M_1	3	σ^2, ϕ, τ^2
M_2	5	$\sigma^2, \phi, \tau^2, \lambda, \gamma$
M_3	4	$\sigma^2, \phi, \tau^2, \alpha$
M_4	6	$\sigma^2, \phi, \tau^2, \gamma, \lambda, \alpha$

Table 1: Number of Parameters and Parameters in Each Model

i) M_1: Stationary Isotropic Model,

$$\boldsymbol{Z}|\boldsymbol{\theta} \sim N_n(\mathbf{0}, \Sigma_1(\boldsymbol{\theta})), \quad \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2),$$

ii) M_2: Stationary Anisotropic Model,

$$\boldsymbol{Z}|\boldsymbol{\theta} \sim N_n(\mathbf{0}, \Sigma_2(\boldsymbol{\theta})), \quad \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \lambda, \gamma),$$

iii) M_3: Nonstationary Point Source Isotropic Model,

$$\boldsymbol{Z}|\boldsymbol{\theta} \sim N_n(\mathbf{0}, \Sigma_3(\boldsymbol{\theta})), \quad \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \alpha),$$

iv) M_4: Nonstationary Point Source Anisotropic Model,

$$\boldsymbol{Z}|\boldsymbol{\theta} \sim N_n(\mathbf{0}, \Sigma_4(\boldsymbol{\theta})), \quad \boldsymbol{\theta} = (\sigma^2, \phi, \tau^2, \gamma, \lambda, \alpha).$$

Note that M_1, M_2, and M_3 are nested under M_4. That is, M_1, M_2, and M_3 are special cases of M_4. Specifically, M_4 reduces to M_1 when $\alpha = 0$ and $A = I$ in the covariance function Σ_4. Also, M_4 reduces to M_2 when $\alpha = 0$ and M_3 when $A = I$. Table 1 summarizes the number of parameters p in each model along with parameters $\boldsymbol{\theta}$ for each model.

4.2 Data Generation Processes

Using the method presented in Cressie (1993, Section 3.6) to simulate point-referenced data, we simulated the spatial process at n locations, $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$, following a multivariate normal distribution with mean vector $E(\boldsymbol{Z}) = \mathbf{0}$, and covariance matrix $Cov(\boldsymbol{Z}) = \Sigma_i, \quad i = 1, \cdots, 4$, as presented in Section 4.1. We used the Cholesky decomposition which allows the covariance matrix,

DGP	α	γ	λ
D_1	0	0	1
D_{21}	0	$\pi/4$	5
D_{22}	0	$\pi/4$	10
D_{31}	5	0	1
D_{32}	10	0	1
D_{41}	5	$\pi/4$	5
D_{42}	5	$\pi/4$	10
D_{43}	10	$\pi/4$	5
D_{44}	10	$\pi/4$	10

Table 2: True Parameter Values for Each Data Generation Process

Σ_i, to be decomposed as the matrix product $\Sigma_i = L_i L_i'$, where L_i is a lower triangular $n \times n$ matrix. Then we simulated $\boldsymbol{Z}$, which satisfies the mean $\mathbf{0}$ and the covariance Σ_i through the relation $\boldsymbol{Z} = L_i \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} = (\epsilon(\boldsymbol{s}_1), \cdots, \epsilon(\boldsymbol{s}_n))'$ and $\epsilon(\boldsymbol{s}_i)$'s are iid with a standard normal distribution. We also simulated irregularly spaced n locations, $\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n$, distributed uniformly on the square $[0, 100]^2$.

Simulated data were generated under 18 different conditions created by varying four factors of interest: the true model ($M = M_1, M_2, M_3, M_4$), the true parameter value for nonstationarity ($\alpha = 5, 10$) and for anisotropy ratio ($\lambda = 5, 10$), and sample size ($n = 50, 100$). From the combinations of the true parameter values in the models, we created nine sets of data as given in Table 2. D_1 was generated from model M_1, D_{21}, D_{22} from model M_2 with different λ, D_{31}, D_{32} from model M_3 with different α, and $D_{41}, D_{42}, D_{43}, D_{44}$ from model M_4 with the combination of different λ and α. We assumed the point source to be located at the origin $\boldsymbol{c} = (0, 0)$ for model M_3 and M_4. These nine data sets were generated with two different sample sizes. 100 data sets were replicated for each of the nine scenarios, and hence a total of 1800 data sets were generated for our simulation study.

4.3 Results

First, we compared the covariance functions of nine scenarios with $n = 50$ given in Table 2 by computing the Frobenius distances between these nine

	Σ_1	Σ_{21}	Σ_{22}	Σ_{31}	Σ_{32}	Σ_{41}	Σ_{42}	Σ_{43}	Σ_{44}
Σ_1	0	8.350	9.745	10.972	11.032	11.120	11.132	11.128	11.133
Σ_{21}		0	2.165	4.810	4.887	4.951	4.979	4.970	4.980
Σ_{22}			0	3.407	3.487	3.413	3.448	3.437	3.450
Σ_{31}				0	0.802	1.721	1.694	1.692	1.689
Σ_{32}					0	1.767	1.741	1.707	1.705
Σ_{41}						0	0.302	0.224	0.312
Σ_{42}							0	0.119	0.076
Σ_{43}								0	0.092
Σ_{44}									0

Table 3: Frobenius Distance between Covariance Functions of Models

covariance functions. Frobenius distance can be used to measure the distance between two matrices and to indicate the difference between these matrices. Suppose $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ are square matrices with the same dimension, then the Frobenius distance between these two matrices is calculated as,

$$F(A, B) = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(a_{ij} - b_{ij})^2}. \tag{12}$$

Thus the closer $F(A, B)$ is to 0, then the more these two matrices A and B are similar. Also notice that $F(A, B) = 0$ if and only if $A = B$.

Table 3 presents the Frobenius distances between the covariance functions of nine scenarios generated using sample size $n = 50$. Here, Σ_i represents the covariance function of D_i, $i = 1, 21, 22, 31, 32, 41, 42, 42, 44$. The Frobenius distances are small between the covariance functions generated from the same model with different parameter values, in particular, between the covariance functions of nonstationary models. Σ_1 seemed very different from all of other covariance functions. Σ_{22} was a little closer than Σ_{21} to the covariance functions of the nonstationary models. That is, the stationary anisotropic model with a large anisotropy ratio ($\lambda = 10$) seems to be closer than the stationary anisotropic model with a small anisotropy ratio ($\lambda = 5$) to the nonstationary model. The distances of the stationary covariance functions from the nonstationary point source isotropic covariance functions are very similar

	$n = 50$				$n = 100$			
	M_1	M_2	M_3	M_4	M_1	M_2	M_3	M_4
AIC	6	10	8	12	6	10	8	12
BIC	11.736	19.560	15.648	23.472	13.816	23.026	18.421	27.631
AICc	6.522	11.364	8.889	13.953	6.250	10.638	8.421	12.903

Table 4: Penalty Given by Each Criterion to Four Models with Two Different Sample Size

to those from the nonstationary point source anisotropic covariance functions. Unlike in the stationary models, whether the covariance function is isotropic or anisotropic seemed not to make much difference in the nonstationary models. The stationary anisotropic model appeared closer to the nonstationary models than to the stationary isotropic model. The distances between nonstationary point source isotropic models and nonstationary point source anisotropic models were the smallest among the distances between different models. Similar comparative features are observed between covariance functions generated using the sample of size $n = 100$.

Each of the nine scenarios was modeled using one of four different models, M_1, M_2, M_3, M_4. Each dataset was thus modeled with one correct model and three incorrect models. Models were fit using the `R` statistical software which uses the `optim` function to maximize the likelihood. AIC, AICc, and BIC were calculated for each dataset using the formula given in (1), (2), and (3), respectively. Table 4 presents the penalty used by each criterion for each model. BIC uses a penalty almost twice as large as that of AIC and AICc. The penalties used by AIC and AICc are not much different. Only the penalty of AIC does not depend on the sample size n.

We examined whether AIC, AICc, and BIC can correctly identify the true underlying model when various spatial models are fit to a particular data set which is generated from one of the models fitted. Results are summarized in Tables 5-8. Each table presents the percentage of times one of the four models is chosen by an information criterion. For example, the first row and second column of Table 4 presents the percentage of times that model M_1 is selected based on AICc when D_1 is fitted. The last row of each table represents the total percentage of times that the true model is not picked by the corresponding criteria, which can be called an 'Error '. In each table, the

Model Fit	DGP: D_1					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1^*	90	99	92	91	100	92
M_2	6	0	5	5	0	4
M_3	4	1	3	4	0	4
M_4	0	0	0	0	0	0
Error	10	1	8	9	0	8

Table 5: Percentage of Correct Decisions When data are generated from M_1

true model is marked by '* ' for convenience.

We expected the true model to be chosen most of the time, when the data are fitted to various models including the true model. Results from our simulation study indicated that AIC, AICc, and BIC performed well for some specific spatial models, however these criteria performed poorly as well for some other spatial models.

Table 5 presents results for D_1 generated from the stationary isotropic model, M_1. All criteria performed well for both $n = 50$ and $n = 100$. Especially BIC performed very well. BIC picked the true model 99% of the time for $n = 50$ and 100% for $n = 100$. AIC and AICc chose the correct model 90% and 92% of the time for $n = 50$ and 91% and 92% of the time for $n = 100$. The performances of AIC and AICc were similar. Note that M_4 was never picked by all criteria. Overall BIC performed better than AIC and AICc in selecting the stationary isotropic model M_1.

Table 6 summarizes the results for D_{21} and D_{22}. D_{21} and D_{22} were generated from the stationary anisotropic model, M_2, with different parameter values for $\lambda = 5$ and $\lambda = 10$, respectively, and with the same parameter value $\gamma = \pi/4$. Each criterion performed similarly in D_{21} and D_{22} except that AIC and AICc picked M_3 more often in D_{22}. All criteria did not perform well when the sample size was $n = 50$. Especially BIC performed poorly. BIC picked the true model, M_2, only 7% of the time in D_{21} and 6% in D_{22}. AIC performed better than AICc for $n = 50$. This is counter to the idea that AICc is designed to perform well for small sample sizes. When $n = 50$, all criteria more often selected M_1 instead of the true model, M_2. All criteria tended to pick the parsimonious model, that is, the simpler model even though the true model is more complex. As sample size increased to 100,

Model Fit	DGP: D_{21}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	50	84	61	20	50	23
M_2^*	32	7	23	71	45	70
M_3	10	8	9	4	5	4
M_4	8	1	7	5	0	3
Error	68	93	77	29	55	30

Model Fit	DGP: D_{22}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	49	83	56	14	41	15
M_2^*	27	6	19	69	50	69
M_3	21	10	23	8	7	8
M_4	3	1	2	9	2	8
Error	73	94	81	31	50	31

Table 6: Percentage of Correct Decisions When data are generated from M_2

the performance of all the criteria improved. All criteria selected the correct model M_2 most often except BIC for D_{21}. AIC and AICc were successful in choosing the correct model and the performances of them were similar. For AIC, the success rate of picking the true model increased from 32% to 71% under D_{21} and from 27% to 69% for D_2. Also, the performances of AICc increased by 47% for D_{21} and 50% for D_{22}. While AIC and AICc performed well with the sample size $n = 100$, BIC still tended to pick parsimonious model, M_1. BIC correctly picked the true model 45% for D_{21} and 50% for D_{22}.

We found that the results from D_{21} and those from D_{22} were similar but slightly different. When $n = 50$, all criteria selected M_3 more often in D_{22} than in D_{21}, and chose M_2 less often in D_{22} than in D_{21} except BIC for $n = 100$. This occurrence makes sense because the covariance function for D_{22} was closer than the covariance function for D_{21} to the covariance functions of M_3. However, based on the Frobenius distance given in Table 3, it still does not make sense to choose M_1 more often than other models.

Model Fit	DGP: D_{31}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	24	49	28	6	17	6
M_2	3	1	2	3	1	3
M_3^*	73	50	70	91	82	91
M_4	0	0	0	0	0	0
Error	27	50	30	9	18	9

Model Fit	DGP: D_{32}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	23	61	35	9	18	9
M_2	4	0	2	4	0	4
M_3^*	73	39	63	87	82	87
M_4	0	0	0	0	0	0
Error	27	61	37	13	18	13

Table 7: Percentage of Correct Decisions When data are generated from M_3

When $n = 100$, all criteria also picked the correct model M_2 less often and picked M_3 and M_4 more often in D_{22} than in D_{21}. We think it is because that the covariance function of D_{22} was closer than that of D_{21} to the covariance functions of M_3 and M_4. BIC picked M_2 more often than M_1 in D_{22}. The reverse occurred in D_{21}. Overall, AIC performed better than AICc and BIC in choosing the stationary anisotropic model M_2.

The results for D_{31} and D_{32} are given in Table 7. D_{31} and D_{32} were generated from the nonstationary point source isotropic model, M_3, with different parameter values for $\theta = 5$ and $\theta = 10$, respectively, and with the same parameter value $\lambda = 1$. Overall, all criteria performed well with the exception of BIC when the sample size was $n = 50$. For $n = 50$, BIC picked both the wrong model, M_1, and the correct model, M_3, with almost the same percentages (49% and 50%, respectively) for D_{31}, while the wrong model, M_1, was picked with a higher percentage (61%) for D_{32}. This indicated that BIC tended to select a simpler model, M_1, than the true model, M_3, when $n = 50$. In contrast, AIC and AICc identified the true model relatively well

for $n = 50$. AIC selected the correct model 73% of the time for both D_{31} and D_{32}, and AICc chose the true model 70% and 63% of the time for D_{31} and D_{32}, respectively. AIC performed better than AICc for the small sample size. For $n = 100$, the performance of all criteria improved. The performances of AIC and AICc were same, and these criteria performed better than BIC. Overall, the error rates decreased with increasing sample size for both D_{31} and D_{32}. All criteria performed better in D_{31} than D_{32}. The performance of each criterion appeared to have a similar pattern under D_{31} and D_{32} except that BIC and AICc picked M_1 more often in D_{31} than in D_{32} for $n = 50$. Note that M_4 was never picked by all criteria given all the values of θ and n considered, even though the covariance function of M_4 was closer than that of M_1 and M_2 to the covariance function of M_3 in terms of the Frobenius distance in Table 3.

Table 8 illustrates the results for D_{41}, D_{42}, D_{43}, and D_{44} which were generated from the nonstationary point source anisotropic model, M_4, with different parameter values of θ and λ as shown in Table 2. As given in Table 8, the performances of the criteria did not vary much across four data sets. All criteria performed very poorly in selecting the true model, M_4, and tended to choose a simpler model than the real model. In particular, BIC did not pick the true model even one time out of 100 replications. Even though AIC performed better than AICc and BIC, AIC did not perform well. AIC only picked the true model less than 5% of the time when $n = 50$ and less than 15% when $n = 100$ for all of four data sets. For $n = 50$, all criteria selected M_1 and M_3 most of the time. BIC picked M_1 more often (more than 70% of the time) than AIC and AICc. AICc picked M_1 more often and selected M_3 less often than AIC. For $n = 100$, AIC and AICc picked M_3 more often, and BIC picked M_1 and M_3 with similar percentage. The performances of AIC and AICc were similar. In this case, it seemed that AIC and AICc made more sense than BIC in model selection for M_4. Selecting M_1 more often than other models seems unreasonable based on the Frobenius distance, because the covariance function of M_1 was much different from that of M_4. The covariance functions of M_2 and M_3 were much closer than that of M_1 to the covariance function of M_4 as shown in Table 3. Overall, AIC performed better than AICc and BIC in selecting M_4.

Model Fit	DGP: D_{41}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	44	76	50	11	47	12
M_2	6	9	4	24	4	22
M_3	46	24	43	57	49	60
M_4^*	4	0	3	8	0	6
Error	96	100	97	92	100	94

Model Fit	DGP: D_{42}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	41	72	49	9	45	9
M_2	8	0	2	21	5	19
M_3	52	28	49	56	50	60
M_4^*	2	0	0	14	0	12
Error	98	100	100	86	100	88

Model Fit	DGP: D_{43}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	34	71	41	14	46	16
M_2	8	1	5	19	4	18
M_3	58	28	54	57	50	59
M_4^*	0	0	0	10	0	7
Error	100	100	100	90	100	93

Model Fit	DGP: D_{44}					
	$n = 50$			$n = 100$		
	AIC	BIC	AICc	AIC	BIC	AICc
M_1	42	75	50	11	49	13
M_2	8	1	4	16	3	12
M_3	49	24	46	64	48	67
M_4^*	1	0	0	9	0	8
Error	99	100	100	91	100	92

Table 8: Percentage of Correct Decisions When data are generated from M_4

5 Discussions and Future Research

We investigated how information criteria such as AIC, AICc, and BIC perform in the spatial model selection problems via simulations. The results are summarized as follows:

- BIC was superior to AIC and AICc when the true model was the stationary isotropic model. When the sample size was large (e.g., $n = 100$), BIC perfectly picked the true model. BIC also performed very well even though the sample size was small (e.g., $n = 50$). AIC and AICc also performed well.

- When the true model was the stationary anisotropic model, all criteria did not perform well for $n = 50$. Especially BIC performed poorly. As n increased to 100, the performance of all criteria improved. AIC and AICc performed well for $n = 100$, however BIC did not perform well even for the large sample size.

- AIC performed better than AICc and BIC for $n = 50$, and both AIC and AICc outperformed BIC for $n = 100$, when the true model was the nonstationary point source isotropic model. BIC picked the stationary isotropic model most often when $n = 50$, however it picked the correct model most of the time when $n = 100$. AIC and AICc performed well for both $n = 50$ and $n = 100$. The error rates for all criteria decreased as sample size increased.

- All criteria performed poorly when the true model was the nonstationary point source anisotropic model. AIC performed better than AICc and BIC. BIC never picked the true model even when the sample size was $n = 100$. In contrast, AIC and AICc picked the true model more often when $n = 100$.

Our results indicate that the performance of the criteria to select the true model generally improved with increase of sample size, despite differences in performance among the criteria. From the results obtained from simulations, we found that the performance of the criteria depends on sample size and model complexity, but not parameter values. Hence, it would be worthwhile to investigate further simulation studies with large sample sizes, e.g., $n = 500$ and $n = 1000$ and other stationary and nonstationary models.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (eds.), *Proceedings of the Second International Symposium on Information Theory.* Budapest: Akademiai Kiado, pp.267-281.

Box, G. E. P. (1979). Some problems of Statistics and everyday life, *Journal of the American Statistical Association*, **74**, 1-4.

Burnham, K. P. and Anderson D. R. (2002). *Model selection and inference: A practical information-theoretic approach*, Springer: New York, 2nd ed.

Cressie, N. A. C. (1993). *Statistics for spatial data.* Wiley: New York, revised edition.

Hoeting, J. A., Davis, R. A., Merton, A. A., and Thomspon, S. E. (2006). Model Selection for Geostatistical Models, *Ecological Applications*, **16:1**, 87-98.

Hughes-Oliver, J. M., Lu, J.-C., Davis, J. C., and Gyurcsik, R. S. (1998) Achieving uniformity in a semiconductor fabrication process using spatial modeling. *Journal of the American Statistical Association*, **93**, 36-45.

Hurvich, C. M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

Hurvich, C. M. and Tsai, C-L. (1990). The impact of model selection on inference in linear regression. *American Statistician*, **44**, 214-217.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22(1)**,79-86.

Lebreton,J-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monograph*, **62**, 67-118.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **12**, 591-612.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*, CRC Press.

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

Spiegelhalter, D. J., Best, N., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of Royal Statistical Society*, Series B, **64**, 583-639.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society*, Series B, **39**, 111-147.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods*, **A7**, 13-26.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences)*, **153**, 12-18. (In Japanese).