Statistica Sinica **28** (2018), 203-228
doi:https://doi.org/10.5705/ss.202016.0368

A GENERALIZED MEASURE OF UNCERTAINTY IN GEOSTATISTICAL MODEL SELECTION

Chun-Shu Chen, Jun Zhu and Tingjin Chu

National Changhua University of Education, University of Wisconsin, Madison and Renmin University of China

Abstract: Model selection and model averaging are essential to regression analysis in environmental studies, but determining which of the two approaches is the more appropriate and under what circumstances remains an active research topic. In this paper, we focus on geostatistical regression models for spatially referenced environmental data. For a general information criterion, we develop a new perturbation-based criterion that measures the uncertainty (or, instability) of spatial model selection, as well as an empirical rule for choosing between model selection and model averaging. Statistical inference based on the proposed model selection instability measure is justified both in theory and via a simulation study. The predictive performance of model selection and model averaging can be quite different when the uncertainty in model selection is relatively large, but the performance becomes more comparable as this uncertainty decreases. For illustration, a precipitation data set in the state of Colorado is analyzed.

Key words and phrases: Data perturbation, generalized degrees of freedom, geostatistics, information criterion, model complexity, spatial prediction.

1. Introduction

For regression analysis of geostatistical data in many environmental studies, the response variable of interest is often observed along with a set of covariates at spatial sampling locations. Selection of a subset of covariates and prediction of the response at unsampled locations are generally based on fitting spatial linear regression models and choosing a suitable subset of covariates using a model selection criterion such as the Akaike's information criterion (AIC) (Akaike (1973)). To assess the fitted models, stochasticity of both the parameter estimates and that of the model selection should be considered. In geostatistics, model selection may involve not only selection of covariates but also determination of the spatial error structure. In this paper, we restrict our attention to the selection of a suitable subset of covariates in a geostatistical model. However, the randomness in the selection of models (or, selection uncertainty) is often ignored in the statistical

inference post model selection (e.g., Breiman (1996)). To mitigate the effect of selection uncertainty, model averaging that pools multiple fitted models is widely used as it can provide a better prediction than a single best model. Although model selection and model averaging have been well studied for the standard linear regression that assumes independent errors (e.g., Shao (1997); Hoeting et al. (1999); Burnham and Anderson (2002); Claeskens and Hjort (2008)), results are far fewer for spatial linear regression in geostatistics. We develop a new approach to geostatistical regression model selection and model averaging for the analysis of spatial data in the environmental sciences.

With model selection (or, covariate selection), generally a best model (or, a best subset of covariates) is selected based on a certain criterion and the more important covariates are identified according to the best model. With model averaging, however, several candidate models are combined based on estimated model weights. Although model averaging tends to give better prediction than model selection, the computational cost is often higher due to the search for suitable weights. It is also more challenging to infer the relationship between the response and the covariates based on an averaged model. Under the Bayesian framework (e.g., Hoeting et al. (1999); Johnson and Hoeting (2011)), the posterior inclusion probability (PIP) of each covariate provides a measure of importance of the covariate in relation to the response, for which a prior specification and Markov chain Monte Carlo (MCMC) for the posterior computation are required. In addition, some Gibbs sampler based methods are also commonly used for Bayesian variable selection such as the stochastic search variable selection (SSVS) algorithm (George and McCulloch (1993)). In this paper, our focus is to investigate the connection between model selection and model averaging in a frequentist framework.

It is in general not clear whether one strategy (i.e., model selection or model averaging) is preferable over the other and under what circumstances. To address this issue, Yuan and Yang (2005) developed a criterion to capture the uncertainty of model selection in standard linear regression with independent errors. Ghosh and Yuan (2009) proposed an L_1-norm criterion that measures the instability of model selection for logistic regression with binary data, as well as an empirical rule to suggest whether model selection or model averaging is preferable. In addition, Efron (2014) used a bootstrap-based method to discuss the stability of an estimator after model selection for independent observations. These results are useful for evaluating model selection and model averaging, but the response variables are assumed to be independent.

In geostatistics, Hoeting et al. (2006) provided some heuristic arguments in spatial linear regression model selection using AIC. A simulation study indicated that if the spatial dependence is ignored, some important covariates may not be selected and hence the prediction errors will be high. Huang and Chen (2007) developed an approximately unbiased estimator of the mean squared prediction error (MSPE) for evaluating different spatial predictors based on generalized degrees of freedom, and derived asymptotic efficiency results for the proposed method, but here the focus was on selection among different predictors obtained by different spatial models for prediction rather than selection of covariates. More recently, Chu, Zhu and Wang (2011) proposed a penalized maximum likelihood estimation (PMLE) and a one-step sparse approximation to simultaneously select covariates and estimate parameters in spatial linear regression models, but quantification of model selection uncertainty was not considered. The uncertainty of model selection has received much attention under the Bayesian approaches (e.g., Clyde and George (2004); Johnson and Hoeting (2011)) and can be measured via posterior inference. To the best of our knowledge, however, foundational questions of how to evaluate the uncertainty of model selection and the connection between model selection and model averaging in geostatistical regression settings have not been adequately addressed under the frequentist framework and will be explored here.

We develop new methodology for geostatistical regression model selection and model averaging in the context of two model selection criteria, a generalized information criterion (GIC) and a conditional generalized information criterion (CGIC). We propose a novel criterion to measure the uncertainty (or, instability) of geostatistical regression model selection based on the selected model and the corresponding predictor. The resulting predictor after model selection and parameter estimation is nonlinear, which makes the problem more challenging to handle than the standard linear regression. Our overall strategy is to develop an index that quantifies the instability in geostatistical regression model selection via a perturbation technique. It simultaneously takes into account the uncertainties of model selection and parameter estimation. By normalizing the instability index, a generalized instability measure is developed which more accurately reflects the complexity of a model fitting procedure. In addition, we establish a theoretical connection between the proposed index of selection instability and the notion of generalized degrees of freedom for geostatistical regression model selection using GIC and CGIC. For practical applications, we further develop an empirical rule that helps to determine whether model selection or model averag-

ing is preferred under GIC and CGIC.

The remainder of this paper is organized as follows. In Section 2, we describe the geostatistical regression model and the corresponding spatial predictor. In Section 3, we derive the properties of various model selection methods for geostatistical regression and some model averaging methods are also introduced. We then develop an index of model selection instability and a generalized instability measure. Further, a theoretical result associated with the proposed methodology is established. In Section 4, we provide an estimation method for the generalized instability measure and an empirical rule for choosing between model selection and model averaging. The results of two simulation scenarios and a weather data example are given in Sections 5 and 6, respectively. We conclude with a discussion in Section 7, and the technical details are given in the Appendix.

2. Geostatistical Regression Model and Spatial Prediction

2.1. Geostatistical regression model

Let $\mathcal{D} \subset \mathbb{R}^2$ be a continuous and bounded study region, and let $\boldsymbol{s}$ be an arbitrary location in $\mathcal{D}$. Suppose there are p covariates at location $\boldsymbol{s}$ that are denoted, together with 1 for the intercept, by $\boldsymbol{x}(\boldsymbol{s}) = (1, x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))'$. A spatial random field $\{S(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ of interest is

$$S(\boldsymbol{s}) = \beta_0 + \sum_{j \in M_0} \beta_j x_j(\boldsymbol{s}) + \eta(\boldsymbol{s}), \tag{2.1}$$

where β_j for $j = 1, \ldots, p$ are regression coefficients, M_0 is the index set of the covariates in the true model, and $\eta(\cdot)$ is a spatial random error process that captures the spatial variation of $S(\cdot)$ and can provide a local adjustment to the mean trend due to unobserved covariates. It is a common practice to assume that $\eta(\cdot)$ follows a Gaussian process with mean zero and covariance function $K(\cdot; \boldsymbol{\theta})$ parameterized by the vector $\boldsymbol{\theta}$. For various covariance functions, see Chapter 4 of Schabenberger and Gotway (2005). Now, the response variable $Z(\boldsymbol{s})$ at location $\boldsymbol{s} \in \mathcal{D}$ is modeled by

$$Z(\boldsymbol{s}) = \beta_0 + \sum_{j \in M_0} \beta_j x_j(\boldsymbol{s}) + \eta(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \tag{2.2}$$

where $\varepsilon(\boldsymbol{s}) \sim N(0, \sigma_\varepsilon^2)$ is a measurement error and is independent of the spatial error process $\eta(\boldsymbol{s})$. We refer to (2.2) as the true geostatistical regression model.

We consider model selection among the p covariates indexed by $\mathcal{P} = \{1, \ldots, p\}$. Let M denote a candidate model as a subset of $\mathcal{P}$, and let $\mathcal{M} \subseteq 2^{\mathcal{P}}$ denote a class of candidate models. Let $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))' \equiv (Z_1, \ldots, Z_n)'$ be the

data observed at n sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, and let $\boldsymbol{X}_M = (\boldsymbol{x}_M(\boldsymbol{s}_1), \ldots, \boldsymbol{x}_M(\boldsymbol{s}_n))'$ be an $n \times (|M|+1)$ design matrix for model M with $|M|$ denoting the number of covariates in M. For a given candidate model $M \in \mathcal{M}$, (2.2) can be rewritten in matrix form as

$$\boldsymbol{Z} = \boldsymbol{X}_M\boldsymbol{\beta}_M + \boldsymbol{\eta} + \boldsymbol{\varepsilon} \sim N(\boldsymbol{X}_M\boldsymbol{\beta}_M, \boldsymbol{\Sigma}_{\boldsymbol{Z}}), \tag{2.3}$$

where $\boldsymbol{\beta}_M$ is the vector of regression coefficients consisting of β_0 and $\{\beta_j : j \in M\}$, $\boldsymbol{\eta} = (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_n))'$ is the vector of spatial random errors with a covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = [K(\boldsymbol{s}_i, \boldsymbol{s}_{i'}; \boldsymbol{\theta})]_{i,i'=1}^n$, $\boldsymbol{\varepsilon} = (\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_n))' \sim N(\boldsymbol{0}, \sigma_\varepsilon^2\boldsymbol{I})$ is the vector of independent measurement errors, and $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) + \sigma_\varepsilon^2\boldsymbol{I}$ is the covariance matrix of the data vector $\boldsymbol{Z}$.

2.2. Spatial prediction

For predicting the spatial random field $\{S(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ based on the data vector $\boldsymbol{Z}$, take $\boldsymbol{S} = (S(\boldsymbol{s}_1), \ldots, S(\boldsymbol{s}_n))'$ at n sampling locations. By (2.1), we have

$$\boldsymbol{S} = \boldsymbol{X}_{M_0}\boldsymbol{\beta}_{M_0} + \boldsymbol{\eta} \sim N(\boldsymbol{X}_{M_0}\boldsymbol{\beta}_{M_0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta})). \tag{2.4}$$

For a given model $M \in \mathcal{M}$ with known $\boldsymbol{\theta}$ and σ_ε^2, the best linear unbiased predictor (BLUP) of $S(\boldsymbol{s})$ at any location $\boldsymbol{s} \in \mathcal{D}$ is obtained by minimizing the MSPE (see, e.g., Schabenberger and Gotway (2005)). The BLUP of $S(\boldsymbol{s})$, indexed by a given model M, is given by

$$\hat{S}_M(\boldsymbol{s}; \boldsymbol{\theta}) = \boldsymbol{x}_M'(\boldsymbol{s})\tilde{\boldsymbol{\beta}}_M + cov(\eta(\boldsymbol{s}), \boldsymbol{\eta})\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}(\boldsymbol{Z} - \boldsymbol{X}_M\tilde{\boldsymbol{\beta}}_M), \tag{2.5}$$

where $\tilde{\boldsymbol{\beta}}_M = (\boldsymbol{X}_M'\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{Z}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ and $cov(\eta(\boldsymbol{s}), \boldsymbol{\eta}) = (K(\boldsymbol{s}, \boldsymbol{s}_1; \boldsymbol{\theta}), \ldots, K(\boldsymbol{s}, \boldsymbol{s}_n; \boldsymbol{\theta}))$. We rewrite the vector of BLUPs $\hat{\boldsymbol{S}}_M(\boldsymbol{\theta}) = (\hat{S}_M(\boldsymbol{s}_1; \boldsymbol{\theta}), \ldots, \hat{S}_M(\boldsymbol{s}_n; \boldsymbol{\theta}))'$ as

$$\hat{\boldsymbol{S}}_M(\boldsymbol{\theta}) = \boldsymbol{H}_M(\boldsymbol{\theta})\boldsymbol{Z}, \tag{2.6}$$

where $\boldsymbol{H}_M(\boldsymbol{\theta}) = \sigma_\varepsilon^2\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M(\boldsymbol{X}_M'\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}$. It follows from (2.6) that $\hat{\boldsymbol{S}}_M(\boldsymbol{\theta})$ is a linear combination of the data vector $\boldsymbol{Z}$, given $\boldsymbol{\theta}$ and σ_ε^2.

In practice, the parameters $\boldsymbol{\theta}$ and σ_ε^2 are unknown but can be estimated by, for example, maximum likelihood (ML), restricted maximum likelihood (REML), or a Bayesian method (e.g., Schabenberger and Gotway (2005)). Here, we use likelihood-based methods to estimate model parameters and in particular, the REML method, because it tends to give less biased estimators than the corresponding ML estimators (McGilchrist (1989); Cressie and Lahiri (1993, 1996)). Let $\hat{\boldsymbol{\theta}}_M$ and $\hat{\sigma}_{\varepsilon,M}^2$ denote the REML estimates under the candidate model (2.3).

We have the estimates of the covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M) + \hat{\sigma}^2_{\varepsilon,M}\boldsymbol{I}$ and the vector of regression coefficients $\hat{\boldsymbol{\beta}}_M = (\boldsymbol{X}'_M\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}'_M\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}\boldsymbol{Z}$. Together with (2.5), an empirical predictor of $S(\boldsymbol{s})$ for any $\boldsymbol{s} \in \mathcal{D}$ is given by

$$\hat{S}_M(\boldsymbol{s};\hat{\boldsymbol{\theta}}_M) = \boldsymbol{x}'_M(\boldsymbol{s})\hat{\boldsymbol{\beta}}_M + \widehat{cov}(\eta(\boldsymbol{s}),\boldsymbol{\eta})\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}(\boldsymbol{Z} - \boldsymbol{X}_M\hat{\boldsymbol{\beta}}_M), \tag{2.7}$$

where $\widehat{cov}(\eta(\boldsymbol{s}),\boldsymbol{\eta}) = (\hat{K}(\boldsymbol{s},\boldsymbol{s}_1;\hat{\boldsymbol{\theta}}_M),\ldots,\hat{K}(\boldsymbol{s},\boldsymbol{s}_n;\hat{\boldsymbol{\theta}}_M))$. Analogous to (2.6), the predictors of $\boldsymbol{S}$ at n sampling locations can be rewritten as

$$\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M) = (\hat{S}_M(\boldsymbol{s}_1;\hat{\boldsymbol{\theta}}_M),\ldots,\hat{S}_M(\boldsymbol{s}_n;\hat{\boldsymbol{\theta}}_M))' = \hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M)\boldsymbol{Z}, \tag{2.8}$$

where

$$\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M) = \hat{\sigma}^2_{\varepsilon,M}\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}\boldsymbol{X}_M(\boldsymbol{X}'_M\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}\boldsymbol{X}_M)^{-1}\boldsymbol{X}'_M\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{Z}}. \tag{2.9}$$

Unlike (2.6), however, the matrix $\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M)$ depends on the data vector $\boldsymbol{Z}$ and hence $\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M)$ in (2.8) is no longer a linear predictor.

3. Index of Selection Instability in Geostatistical Regression

3.1. Model selection via GIC and CGIC

Under the candidate model (2.3), we consider the generalized information criterion (GIC),

$$\mathrm{GIC}_\lambda(M) = -2\ell_M\big(\hat{\boldsymbol{\beta}}_M,\hat{\boldsymbol{\theta}}_M,\hat{\sigma}^2_{\varepsilon,M};\boldsymbol{Z}\big) + \lambda(|M| + |\boldsymbol{\theta}| + 2), \tag{3.1}$$

where $\lambda > 0$ is a penalty parameter, $|M|$ is the number of covariates in model M, $|\boldsymbol{\theta}|$ is the number of unknown parameters in $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$, $\ell_M(\cdot)$ is the log-likelihood function of $\boldsymbol{Z}$ based on model M, and $\hat{\boldsymbol{\beta}}_M$, $\hat{\boldsymbol{\theta}}_M$, and $\hat{\sigma}^2_{\varepsilon,M}$ are the REML estimates. The GIC in (3.1) includes Akaike's information criterion (AIC) with $\lambda = 2$ (Akaike (1973)), the Bayesian information criterion (BIC) with $\lambda = \log(n)$ (Schwarz (1978)), the corrected AIC (AICc) criterion with $\lambda = 2n/(n - |M| - |\boldsymbol{\theta}| - 2)$ (Hurvich and Tsai (1989)), and the risk inflation criterion (RIC) with $\lambda = 2\log(p)$ (Foster and George (1994)). For a given λ, the model that has the smallest value of GIC_λ is selected as the best model and is denoted by

$$\hat{M}(\lambda) = \arg\min_{M\in\mathcal{M}} \mathrm{GIC}_\lambda(M). \tag{3.2}$$

By (2.8) and (2.9), the corresponding predictor of $\boldsymbol{S}$ is

$$\hat{\boldsymbol{S}}_{\hat{M}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{M}(\lambda)}) = \hat{\boldsymbol{H}}_{\hat{M}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{M}(\lambda)})\boldsymbol{Z}. \tag{3.3}$$

In addition, we consider a conditional generalized information criterion (CGIC) given by

$$\mathrm{CGIC}_\lambda(M) = n^{-1}\left\{\|\boldsymbol{Z} - \hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M)\|^2 + \lambda\hat{\sigma}^2_\varepsilon\ \mathrm{tr}\left(\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M)\right)\right\}, \tag{3.4}$$

where $\lambda > 0$ is a penalty parameter and $\hat{\sigma}_\varepsilon^2$ is an estimate of σ_ε^2 invariant to the model choice obtained by, for example, REML based on the full model. The CGIC in (3.4) includes the conditional Akaike's information criterion (CAIC) with $\lambda = 2$ and the conditional BIC (CBIC) criterion with $\lambda = \log(n)$ as special cases (e.g., Vaida and Blanchard (2005); Chen and Huang (2012)). For a given λ, the selected model based on CGIC_λ is denoted by

$$\hat{M}_c(\lambda) = \arg\min_{M \in \mathcal{M}} \mathrm{CGIC}_\lambda(M) \tag{3.5}$$

and the corresponding predictor of $\boldsymbol{S}$ is

$$\hat{\boldsymbol{S}}_{\hat{M}_c(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{M}_c(\lambda)}) = \hat{\boldsymbol{H}}_{\hat{M}_c(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{M}_c(\lambda)})\boldsymbol{Z}. \tag{3.6}$$

Compared with the GIC in (3.1), the CGIC in (3.4) not only considers the size of regression term (i.e., $|M|$), but also the complexity of spatial dependence in the model selection procedure. The proof of the following result is given in the Appendix.

Proposition 1. *Under* (2.3), (2.8), *and* (2.9), *if* $|M|$ *is the number of covariates in model* M *and* $\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) = \boldsymbol{X}_M(\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}$ *with* $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M) + \hat{\sigma}_{\varepsilon,M}^2\boldsymbol{I}$, *then, for any* $\lambda > 0$ *and* $M \in \mathcal{M}$, (3.4) *can be rewritten as*

$$\begin{aligned} CGIC_\lambda(M) = n^{-1}[\|\boldsymbol{Z} - \hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M)\|^2 + \lambda\hat{\sigma}_\varepsilon^2|M| \\ + \lambda\hat{\sigma}_\varepsilon^2\mathrm{tr}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\{\boldsymbol{I} - \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\})]. \end{aligned} \tag{3.7}$$

The third term on the right side of (3.7) reflects the complexity of spatial dependence in the model selection procedure. Hence, the CGIC is expected to be more suitable than the GIC in geostatistical regression model selection, as will be demonstrated by a simulation study in Section 5.

3.2. Model averaging via GIC and CGIC

Model averaging combines several predictors $\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M)$ obtained from candidate models for $M \in \mathcal{M}$ based on estimated model weights. Two approaches to estimating the model weights are frequentist model averaging (FMA) and Bayesian model averaging (BMA). We focus on using the FMA. Readers interested in BMA may refer to Raftery, Madigan and Hoeting (1997) and Hoeting et al. (1999) for details. We first review a commonly used FMA technique based on GIC and CGIC (e.g., Burnham and Anderson (2002); Claeskens and Hjort (2008)). The idea is to consider all the candidate models in $\mathcal{M}$ for averaging, where the weight of each candidate model M is determined by

$$\hat{w}_\lambda(M) = \frac{\exp\{-(1/2)\mathrm{GIC}_\lambda(M)\}}{\sum_{M^*\in\mathcal{M}}\exp\{-(1/2)\mathrm{GIC}_\lambda(M^*)\}} \tag{3.8}$$

for $\mathrm{GIC}_\lambda(M)$, or

$$\hat{w}^c_\lambda(M) = \frac{\exp\{-(1/2)\mathrm{CGIC}_\lambda(M)\}}{\sum_{M^*\in\mathcal{M}}\exp\{-(1/2)\mathrm{CGIC}_\lambda(M^*)\}} \tag{3.9}$$

for $\mathrm{CGIC}_\lambda(M)$. For a given $\lambda > 0$, the model averaging predictors of $\boldsymbol{S}$ based on $\mathrm{GIC}_\lambda(M)$ and $\mathrm{CGIC}_\lambda(M)$ are obtained, and are respectively denoted as

$$\hat{\boldsymbol{S}}_\lambda = \sum_{M\in\mathcal{M}} \hat{w}_\lambda(M)\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M), \tag{3.10}$$

and

$$\hat{\boldsymbol{S}}^c_\lambda = \sum_{M\in\mathcal{M}} \hat{w}^c_\lambda(M)\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M), \tag{3.11}$$

where $\hat{\boldsymbol{S}}_M(\hat{\boldsymbol{\theta}}_M)$ is given in (2.8). See Burnham and Anderson (2002) and Claeskens and Hjort (2008) for a comprehensive review of model averaging.

3.3. Generalized instability measure

If a model selection procedure is unstable, model averaging strategies can be considered in order to make more accurate predictions. Otherwise, it is relatively easy to find a best model according to some criterion, and then predictions will work well based on the selected model. Therefore, a measure of the instability associated with a model selection procedure is a critical issue when deciding model selection or model averaging.

We develop new measures of the instability associated with model selection based on GIC and CGIC. Let $\boldsymbol{e}_i$ for $i = 1, \ldots, n$ be the ith column of the $n \times n$ identity matrix. For a given penalty parameter $\lambda > 0$ (or, a given model selection criterion), we define an index of selection instability (ISI) as

$$\mathrm{ISI}(\lambda) = E\left(\lim_{\delta\to 0}\delta^{-1}\sum_{i=1}^{n}\left|\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i) - \hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\right|\right), \tag{3.12}$$

where $\hat{\gamma}(\lambda)$ denotes a selected model obtained by GIC_λ or CGIC_λ according to the data vector $\boldsymbol{Z}$. Here, $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i)$ and $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})$ are the predictors of $S(\boldsymbol{s}_i)$ based on the model $\hat{\gamma}(\lambda)$ applied to $\boldsymbol{Z}+\delta\boldsymbol{e}_i$ and $\boldsymbol{Z}$, respectively. It can be shown that $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}) = \hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ and thus, $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})$ is an alternative notation of $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$, used here to emphasize that it relies on the data vector $\boldsymbol{Z}$.

The ISI is an L_1-norm criterion and can be applied to assess the instability of $\mathrm{GIC}_\lambda(M)$ of (3.1) and $\mathrm{CGIC}_\lambda(M)$ of (3.4). If a model selection procedure

is unstable, a minor perturbation in $\boldsymbol{Z}$ has a high chance to select a very different model, say $\hat{\gamma}^*(\lambda)$ and $\hat{\gamma}^*(\lambda) \neq \hat{\gamma}(\lambda)$. As a result, the difference between $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i; \boldsymbol{Z} + \delta \boldsymbol{e}_i)$ and $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i; \boldsymbol{Z})$ under the same model $\hat{\gamma}(\lambda)$ is expected to be large. Thus, the predictors based on the selected model tend to have large variances and the ISI value of (3.12) is expected to be larger. Unlike the standard regression models with independent responses (Ghosh and Yuan (2009)), our focus is on geostatistical regression models with spatially dependent responses and the predictors in (3.12) are nonlinear after model selection and parameter estimation.

The following proposition provides an alternative expression for ISI(λ); the proof is in the Appendix.

Proposition 2. *Under* (2.3), (2.4), (2.8), *and* (2.9), *if* $\hat{\gamma}(\lambda)$ *is obtained from* $GIC_\lambda(M)$ *of* (3.1) *or* $CGIC_\lambda(M)$ *of* (3.4) *with corresponding predictors* $\hat{\boldsymbol{S}}_{\hat{\gamma}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}) = (\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_1; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}), \ldots, \hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_n; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}))'$ *of* $\boldsymbol{S}$, *and if* $\sum_{i=1}^n E(|\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})||\boldsymbol{S}) < \infty$ *almost surely, then,* $ISI(\lambda)$ *of* (3.12) *is given by*

$$ISI(\lambda) = \sum_{i=1}^{n} E\left(\frac{\partial}{\partial S(\boldsymbol{s}_i)} E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i; \hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}) \big| \boldsymbol{S}\big) \right) . \tag{3.13}$$

With $\lambda > 0$, ISI(λ) at (3.13) can be interpreted as the expected sum of sensitivities of the spatial predictor $\hat{S}_{\hat{\gamma}(\lambda)}(\cdot)$ with respect to the underlying process $S(\cdot)$. Thus, a larger ISI value indicates that the corresponding model selection criterion has a higher selection instability. Although it is consistent with the interpretation of the definition of ISI(λ) at (3.12), (3.13) gives a more intuitive interpretation regarding the instability of a model selection criterion.

The ISI(λ) at (3.13) is akin to a generalized degrees of freedom which can be used to measure the complexity of a model fitting procedure (e.g., Ye (1998); Huang and Chen (2007)). In the absence of spatial dependence, ISI(λ) reduces to the generalized degrees of freedom (GDF) in Ye (1998).

For fairer comparison among various model selection criteria, we normalize the ISI(λ) and define a generalized instability measure (GIM) as

$$\text{GIM}(\lambda) = \frac{\text{ISI}(\lambda)}{|\hat{\gamma}(\lambda)| + |\boldsymbol{\theta}| + 2} , \tag{3.14}$$

where $\hat{\gamma}(\lambda)$ is the model selected by GIC_λ or CGIC_λ, $|\hat{\gamma}(\lambda)|$ is the number of covariates in model $\hat{\gamma}(\lambda)$, $|\boldsymbol{\theta}|$ is the number of unknown parameters in $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$, and the additional 2 is for the intercept β_0 and the noise variance σ_ε^2. We use GIM(λ) to measure the instability of a model selection criterion under geostatistical regression settings.

4. Practical Considerations

4.1. Estimation of the generalized instability measure

In general, ISI(λ) and GIM(λ) are unknown and need to be estimated. For this purpose, we obtain an approximately unbiased estimator $\widehat{\text{ISI}}(\lambda)$ of ISI(λ) based on a data perturbation technique (Huang and Chen (2007)). Here, we use $\text{GIC}_\lambda(M)$ of (3.1) to illustrate how to estimate ISI(λ), a similar procedure can be applied to $\text{CGIC}_\lambda(M)$ of (3.4).

We consider model selection among the p covariates indexed by $\mathcal{P} = \{1, \ldots, p\}$. Let M denote a candidate model as a subset of $\mathcal{P}$ and let $\mathcal{M} \subseteq 2^{\mathcal{P}}$ denote a class of candidate models. For a given data vector $\boldsymbol{Z}$ and a selection criterion $\text{GIC}_\lambda(M)$ with $\lambda > 0$ and $M \in \mathcal{M}$, the estimation procedure of ISI(λ) comprises four steps.

1. Based on (3.1) and (3.2), select a model $\hat{M}(\lambda)$ with $|\hat{M}(\lambda)|$ covariates from $\mathcal{M}$.

2. Generate a set of perturbed data vectors $\boldsymbol{Z}^{*(t)} = (Z_1^{*(t)}, \ldots, Z_n^{*(t)})' = \boldsymbol{Z} + \tau\boldsymbol{\xi}^{(t)}$ for $t = 1, \ldots, T$, with $\boldsymbol{\xi}^{(t)} \sim N(\boldsymbol{0}, \hat{\sigma}_\varepsilon^2 \boldsymbol{I})$ independent of $\boldsymbol{Z}$ and $\tau \in (0, 1]$ the size of perturbation.

3. Based on (3.1) and (3.2), select a model $\hat{M}^{*(t)}(\lambda)$ from $\mathcal{M}$ for each perturbed data vector $\boldsymbol{Z}^{*(t)}; t = 1, \ldots, T$, where the corresponding predictor of $\boldsymbol{S}$ is denoted as $\hat{\boldsymbol{S}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}(\hat{\boldsymbol{\theta}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}) = (\hat{S}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}(\boldsymbol{s}_1; \hat{\boldsymbol{\theta}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}), \ldots, \hat{S}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}(\boldsymbol{s}_n; \hat{\boldsymbol{\theta}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}))'$.

4. With $\bar{S}^*(\boldsymbol{s}_i) \equiv T^{-1} \sum_{t=1}^T \hat{S}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}(\boldsymbol{s}_i; \hat{\boldsymbol{\theta}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)})$ and $\bar{Z}_i^* \equiv T^{-1} \sum_{t=1}^T Z_i^{*(t)}$ for $i = 1, \ldots, n$, approximate ISI(λ) by

$$\widehat{\text{ISI}}(\lambda) = \frac{1}{(T-1)\tau^2\hat{\sigma}_\varepsilon^2} \sum_{i=1}^n \sum_{t=1}^T (\hat{S}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}(\boldsymbol{s}_i; \hat{\boldsymbol{\theta}}^{*(t)}_{\hat{M}^{*(t)}(\lambda)}) - \bar{S}^*(\boldsymbol{s}_i))(Z_i^{*(t)} - \bar{Z}_i^*).$$

From $\widehat{\text{ISI}}(\lambda)$, an estimator $\widehat{\text{GIM}}(\lambda)$ of GIM(λ) for GIC_λ is then

$$\widehat{\text{GIM}}(\lambda) = \frac{\widehat{\text{ISI}}(\lambda)}{|\hat{M}(\lambda)| + |\boldsymbol{\theta}| + 2} \ . \tag{4.1}$$

Henceforth, we use $\widehat{\text{GIM}}(\lambda)$ to measure the instability of a model selection criterion when fitting a geostatistical regression model.

4.2. An empirical rule

We propose an empirical rule based on GIM(λ) to roughly estimate the size

of the underlying true model. For $j = 1, \ldots, p$, let $\mathcal{P}_j$ be the class of all subsets of size j that can be selected from $\mathcal{P} = \{1, \ldots, p\}$. Let $\hat{\gamma}_j(\lambda)$ be the best subset among $\mathcal{P}_j$ selected by GIC_λ or CGIC_λ. Let $\mathcal{P}_0$ be the subset that contains the intercept-only model. Based on (3.14), take the instability measure for $\mathcal{P}_j$ as

$$\text{GIM}_j(\lambda) = \frac{\text{ISI}_j(\lambda)}{j + |\boldsymbol{\theta}| + 2}\ , \tag{4.2}$$

where $\text{ISI}_j(\lambda)$ is akin to $\text{ISI}(\lambda)$ at (3.12) but selects models from $\mathcal{P}_j$. The estimate $\widehat{\text{GIM}}_j(\lambda)$ of $\text{GIM}_j(\lambda)$ can be obtained by the estimation procedure of $\text{GIM}(\lambda)$, where $\mathcal{M}$ is replaced with $\mathcal{P}_j$. Intuitively, if the true model $M_0 \in \mathcal{P}_{j^*}$, the value of $\text{GIM}_j(\lambda)$ is expected to decrease as j increases for $0 \leq j < j^*$ and is expected to be stable for $j^* \leq j \leq p$. To see the pattern of change in $\big\{\widehat{\text{GIM}}_j(\lambda) : j = 0, 1, \ldots, p\big\}$ more clearly, we define a relative difference of generalized instability measures between $\mathcal{P}_j$ and $\mathcal{P}_{j-1}$ as

$$\text{ReGIM}_j(\lambda) = \big|\text{GIM}_j(\lambda) - \text{GIM}_{j-1}(\lambda)\big|; \quad j = 1, \ldots, p. \tag{4.3}$$

In practice, we judge the size of the underlying true model according to the pattern of $\big\{\widehat{\text{ReGIM}}_j(\lambda) : j = 1, \ldots, p\big\}$.

5. Simulation Study

We conducted two simulation scenarios to evaluate the performance of the generalized instability measure $\widehat{\text{GIM}}(\lambda)$ in (4.1) for geostatistical regression model selection and model averaging.

We considered an isotropic and stationary process for the spatial random error process $\eta(\cdot)$ with a Matérn covariance function $K(\boldsymbol{s}_A, \boldsymbol{s}_B; \boldsymbol{\theta}) \equiv \sigma_\eta^2 \rho(\boldsymbol{s}_A, \boldsymbol{s}_B; a, \nu)$ and $\boldsymbol{\theta} \equiv (\sigma_\eta^2, a, \nu)'$, where $\rho(\boldsymbol{s}_A, \boldsymbol{s}_B; a, \nu)$ is a Matérn correlation function (Matérn (2013)) defined by

$$\rho(\boldsymbol{s}_A, \boldsymbol{s}_B; a, \nu) = \frac{\|\boldsymbol{s}_A - \boldsymbol{s}_B\|^\nu}{\Gamma(\nu) a^\nu 2^{\nu-1}}\, \mathcal{K}_\nu(a^{-1}\|\boldsymbol{s}_A - \boldsymbol{s}_B\|). \tag{5.1}$$

In (5.1), $\|\boldsymbol{s}_A - \boldsymbol{s}_B\| > 0$ is Euclidean distance, $\mathcal{K}_\nu(\cdot)$ is a modified Bessel function of the second kind with order $\nu > 0$, ν is a smoothness parameter, $a > 0$ is a range parameter, and σ_η^2 is the variance of $\eta(\cdot)$. A larger ν value indicates a smoother spatial process, while a larger a value indicates a stronger spatial dependence.

We considered simulation scenarios I (strong spatial dependence) and II (weak spatial dependence). All results were based on $B = 200$ simulation replicates.

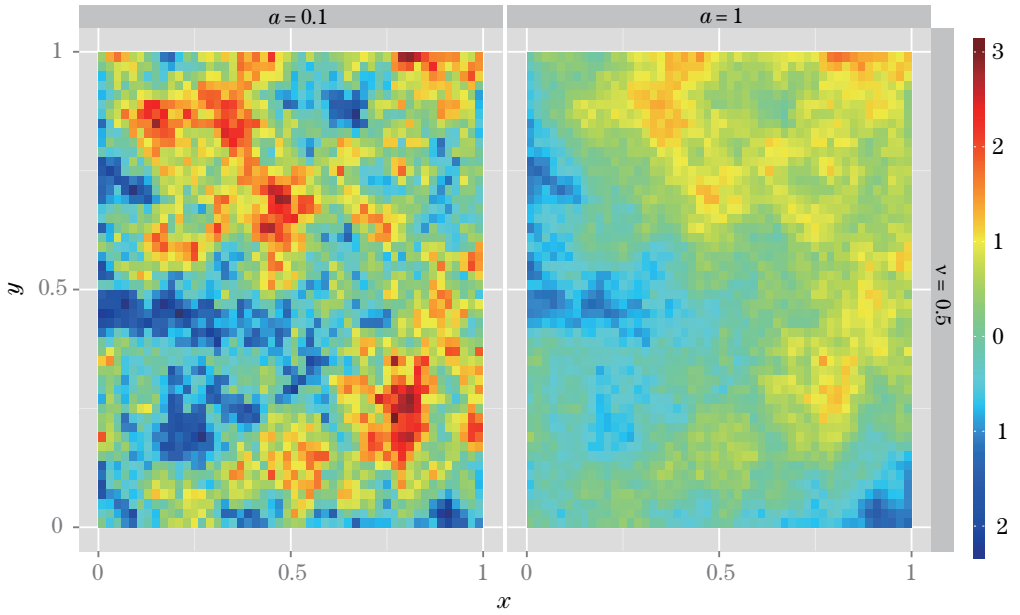

Figure 1. Realizations of $\eta(\cdot)$ for $(\sigma_\eta^2, a, \nu) = (1, 0.1, 0.5)$ (left) and $(\sigma_\eta^2, a, \nu) = (1, 1, 0.5)$ (right).

5.1. Simulation scenario I

Under the true model (2.2), we took $\mathcal{D} \equiv [0,1]^2$ in $\mathbb{R}^2$ and $N = 100 \times 100 = 10,000$ regular grid points, with the coordinate of the ith grid point denoted by $\boldsymbol{s}_i = (s_{i1}, s_{i2})'$ for $i = 1, \ldots, N$. In our simulation scenario, a total of $p = 9$ covariates $\{x_1(\boldsymbol{s}), \ldots, x_9(\boldsymbol{s}); \boldsymbol{s} \in \mathcal{D}\}$ were considered and each covariate was independently generated from the standard Gaussian. The spatial random error process $\eta(\cdot)$ was a zero-mean Gaussian stationary process, where the covariance matrix was the exponential covariance function (i.e., $\nu = 0.5$ in (5.1)) with $\boldsymbol{\theta} = (1, 1, 0.5)'$. The right side of Figure 1 shows a realization of $\eta(\cdot)$ that corresponds to a stronger spatial dependence. The measurement error variance $\sigma_\varepsilon^2 = 1$ in (2.2) was assumed known throughout simulation scenario I. We considered five different M_0 as the underlying true models in the forms of $\beta_0 + \sum_{j=1}^{k} \beta_j x_j(\boldsymbol{s})$ for $k = 1, 3, 5, 7$, and 9. For each k, the regression coefficients were set to be $\beta_0 = 1$, $\beta_1 = \cdots = \beta_k = (3/k)^{1/2}$, and $\beta_{k+1} = \cdots = \beta_p = 0$, so that the signal-to-noise ratio (SNR) was kept at 4, where the SNR is defined as the ratio of the variance of the signal $S(\cdot)$ to the noise variance σ_ε^2.

In each simulated data set, a sample size of $n = 50$ was drawn from the $10,000$ grid points based on a simple random sampling scheme. For each of

$k = 1, 3, 5, 7$, and 9, we applied six model selection criteria, AIC, BIC, AICc, RIC, CAIC, and CBIC, to select models among $\mathcal{M}$, where $\mathcal{M} = 2^{\{1,\ldots,9\}}$ consists of all the possible combinations of covariates $x_1, x_2, \ldots, x_9$, with $M = \emptyset \in \mathcal{M}$ representing the intercept-only model. Throughout simulation scenario I, the model parameters were estimated by REML.

To assess the instability of model selection based on AIC, BIC, AICc, RIC, CAIC, and CBIC, the corresponding values of $\widehat{\mathrm{GIM}}(\lambda)$ in (4.1) were computed. The perturbed sample size was set to $T = 50$ and the perturbation size was set to $\tau = 0.1$ for computing $\widehat{\mathrm{ISI}}(\lambda)$.

To compare with six model selection criteria, predictions based on model averaging methods were also conducted. For each of $k = 1, 3, 5, 7$, and 9, the model averaging methods based on (3.8)–(3.11) are referred to as AIC-MA, BIC-MA, AICc-MA, RIC-MA, CAIC-MA, and CBIC-MA.

5.2. Simulation result I

We compared the prediction performance of various model selection and model averaging methods by examining an average squared prediction error (ASPE):

$$\mathrm{ASPE} = B^{-1} \sum_{b=1}^{B} \left[n^{-1} \sum_{i=1}^{n} \left\{ \hat{S}^{(b)}(\boldsymbol{s}_i) - S^{(b)}(\boldsymbol{s}_i) \right\}^2 \right], \tag{5.2}$$

where $\hat{S}^{(b)}(\boldsymbol{s})$ is a generic predictor of $S^{(b)}(\boldsymbol{s})$ corresponding to (3.3), (3.6), (3.10), or (3.11) for the bth simulated data set and $S^{(b)}(\boldsymbol{s})$ is an underlying random variable of interest at the location $\boldsymbol{s}$ for the bth simulation replicate. In addition, we examined the performance of variable selection for GIC_λ and CGIC_λ under five true models. Table 1 shows the ASPE values for the six model selection criteria, AIC, BIC, AICc, RIC, CAIC, and CBIC, and the six model averaging methods, AIC-MA, BIC-MA, AICc-MA, RIC-MA, CAIC-MA, and CBIC-MA, for $k = 1, 3, 5, 7$, and 9. The corresponding variable selection results for the model selection criteria are given in Table 2. For ease of comparison, the ASPE values in Table 1 are also plotted in Figure 2(a), where we omit the results of AICc, AICc-MA, RIC, and RIC-MA because the patterns are similar.

Table 1 and Figure 2(a) indicate that the CGIC_λ performs well in most cases, as it is closely related to the MSPE. For example, the CAIC is an unbiased estimator of the MSPE plus the noise variance when the model parameters are known (Vaida and Blanchard (2005)). Comparing AIC versus AIC-MA, BIC versus BIC-MA, etc. in terms of the ASPE values, we find that $k = 1, 3$, or

Table 1. Average squared prediction errors of six model selection criteria (AIC, BIC, AICc, RIC, CAIC, CBIC) and six model averaging methods (AIC-MA, BIC-MA, AICc-MA, RIC-MA, CAIC-MA, CBIC-MA) under five true models (k) based on 200 simulation replicates for simulation scenario I. The values in parentheses are the corresponding standard errors.

Criterion	k				
	1	3	5	7	9
AIC	0.5357	0.5318	0.5264	0.5201	0.5172
	(0.0193)	(0.0189)	(0.0188)	(0.0184)	(0.0184)
AIC-MA	0.4598	0.4739	0.4814	0.4944	0.5130
	(0.0184)	(0.0180)	(0.0174)	(0.0161)	(0.0150)
BIC	0.5044	0.5055	0.5029	0.5082	0.5379
	(0.0216)	(0.0205)	(0.0200)	(0.0188)	(0.0187)
BIC-MA	0.4352	0.4568	0.4693	0.4924	0.5267
	(0.0190)	(0.0184)	(0.0175)	(0.0151)	(0.0134)
AICc	0.5284	0.5241	0.5057	0.5123	0.5325
	(0.0204)	(0.0203)	(0.0197)	(0.0190)	(0.0186)
AICc-MA	0.4461	0.4627	0.4722	0.4912	0.5205
	(0.0187)	(0.0182)	(0.0175)	(0.0154)	(0.0137)
RIC	0.4876	0.4993	0.4842	0.5079	0.5488
	(0.0217)	(0.0207)	(0.0191)	(0.0188)	(0.0182)
RIC-MA	0.4303	0.4540	0.4677	0.4945	0.5359
	(0.0192)	(0.0185)	(0.0175)	(0.0148)	(0.0131)
CAIC	0.3866	0.4086	0.4263	0.4300	0.4580
	(0.0155)	(0.0158)	(0.0141)	(0.0111)	(0.0114)
CAIC-MA	0.3650	0.3749	0.3974	0.4198	0.4660
	(0.0159)	(0.0126)	(0.0118)	(0.0094)	(0.0101)
CBIC	0.3720	0.4416	0.4482	0.4699	0.5508
	(0.0164)	(0.0199)	(0.0169)	(0.0141)	(0.0145)
CBIC-MA	0.3538	0.4092	0.4190	0.4597	0.5303
	(0.0162)	(0.0175)	(0.0136)	(0.0116)	(0.0111)

5 with fewer covariates in our simulation, the ASPE values of model selection and model averaging give quite different prediction results. For example, AIC with an average ASPE value 0.5357 and a standard error 0.0193 is significantly different to AIC-MA which has an average ASPE value 0.4598 with a standard error 0.0184 (see, e.g., $k = 1$ in Table 1). In general, model averaging outperforms model selection. However, the prediction results of model selection and model averaging are more comparable for $k = 7$ and 9 here.

In terms of the rate of selecting the true model (Table 2), a model selection criterion with a larger penalty parameter (e.g., BIC, RIC, and CBIC) penalizes more for a model with more covariates and hence tends to have a higher rate of

Table 2. Frequencies of the number of selected covariates for six information criteria (AIC, BIC, AICc, RIC, CAIC, CBIC) under five true models (k) based on 200 simulation replicates for simulation scenario I. The symbol "$*$" indicates that the true model is selected.

k	Criterion	Number of covariates										Average number of selected covariates
		0	1	2	3	4	5	6	7	8	9	
1	AIC	0	0*	0	0	3	12	12	24	34	115	8.10
	BIC	0	48*	79	55	13	5	0	0	0	0	2.24
	AICc	0	1*	19	57	77	43	3	0	0	0	3.76
	RIC	0	76*	85	32	6	1	0	0	0	0	1.86
	CAIC	0	28*	55	60	31	17	4	4	1	0	2.94
	CBIC	0	78*	63	24	15	11	3	4	2	0	2.27
3	AIC	0	0	0	0*	0	2	13	22	35	128	8.37
	BIC	0	0	0	70*	84	43	2	1	0	0	3.90
	AICc	0	0	0	10*	72	82	31	5	0	0	4.75
	RIC	0	0	0	106*	67	26	1	0	0	0	3.61
	CAIC	0	0	2	45*	73	52	21	6	1	0	4.34
	CBIC	0	0	12	98*	61	18	5	4	2	0	3.63
5	AIC	0	0	0	0	0	1*	4	11	44	140	8.59
	BIC	0	0	0	0	0	103*	77	19	1	0	5.59
	AICc	0	0	0	0	0	58*	101	36	5	0	5.94
	RIC	0	0	0	0	1	129*	60	10	0	0	5.40
	CAIC	0	0	1	0	15	79*	72	29	4	0	5.62
	CBIC	0	0	0	3	31	113*	43	8	2	0	5.14
7	AIC	0	0	0	0	0	0	0	4*	36	160	8.78
	BIC	0	0	0	0	0	1	6	140*	51	2	7.24
	AICc	0	0	0	0	0	0	2	118*	75	5	7.42
	RIC	0	0	0	0	0	1	11	151*	35	2	7.13
	CAIC	0	0	0	0	1	4	32	106*	52	5	7.10
	CBIC	0	0	0	1	2	12	50	110*	24	1	6.71
9	AIC	0	0	0	0	0	0	0	0	2	198*	8.99
	BIC	0	0	0	0	0	0	3	5	30	162*	8.76
	AICc	0	0	0	0	0	0	0	2	31	167*	8.83
	RIC	0	0	0	0	0	0	4	9	46	141*	8.62
	CAIC	0	0	0	0	0	1	6	16	56	121*	8.45
	CBIC	0	0	0	0	1	8	17	31	72	71*	7.89

selecting the true model when the underlying true model has fewer covariates, and vice versa. By Proposition 1, CGIC_λ considers the complexity of a spatial dependence in its selection procedure. As expected, CGIC_λ outperforms the corresponding GIC_λ. Somewhat surprisingly, the RIC criterion tends to have a relatively high rate of selecting the true model. Besides the conditional informa-

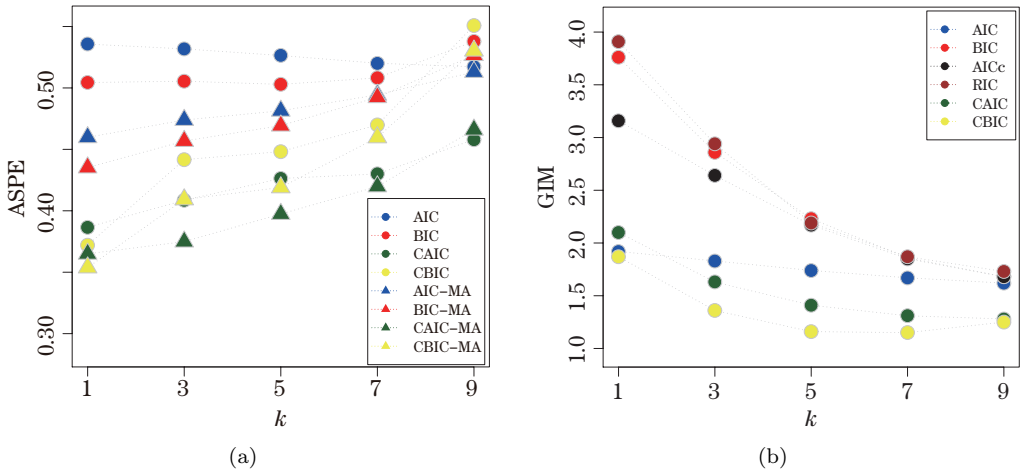

Figure 2. (a) Average squared prediction error (ASPE) versus five true models (k) for four information criteria (AIC, BIC, CAIC, CBIC) and four model averaging methods (AIC-MA, BIC-MA, CAIC-MA, CBIC-MA) based on 200 simulation replicates under simulation scenario I; (b) Average value of generalized instability measure (GIM) versus five true models (k) for six information criteria (AIC, BIC, AICc, RIC, CAIC, CBIC) based on 200 simulation replicates under simulation scenario I.

tion criteria and our empirical rule, RIC is a suitable criterion for determining the size of the underlying true model.

The left panel of Table 3 shows the generalized instability measure under the AIC, BIC, AICc, RIC, CAIC, and CBIC criteria by computing $\widehat{\mathrm{GIM}}(\lambda)$ of (4.1), and the corresponding results are plotted in Figure 2(b). CAIC and CBIC tend to have smaller GIM values than AIC, BIC, AICc, and RIC under geostatistical regression model selection; this agrees with Proposition 1. The conditional information criteria are more stable and thus are more suitable in geostatistical regression model selection, as shown in Tables 1 and 2. From a prediction point of view and for a given model selection criterion, Figures 2(a) and 2(b) show that model selection and model averaging can be quite different when model selection uncertainty is relatively large, but tend to be more similar as model selection uncertainty decreases.

We performed a sensitivity analysis to evaluate the effect of the perturbation size τ on the computation of $\widehat{\mathrm{GIM}}(\lambda)$. The average values of $\widehat{\mathrm{GIM}}(\lambda)$ based on $k = 5$ and 200 simulation replicates are shown in the right panel of Table 3, along with standard errors. The numerical results indicate that the computation of $\widehat{\mathrm{GIM}}(\lambda)$ is not sensitive to the choice of τ. Huang and Chen (2007) and Chen, Yang and Li (2014) also showed that model selection and weight selection

Table 3. Average values of generalized instability measure (GIM) of six information criteria (AIC, BIC, AICc, RIC, CAIC, CBIC) under five true models (k) (left panel) for simulation scenario I and sensitivity analysis of the proposed generalized instability measure (GIM) with respect to various perturbation sizes τ for $k = 5$ (right panel) based on 200 simulation replicates under simulation scenario I. The values in parentheses are the corresponding standard errors.

Criterion	k					τ				
	1	3	5	7	9	0.1	0.3	0.5	0.7	0.9
AIC	1.92	1.83	1.74	1.67	1.62	1.74	1.69	1.68	1.70	1.70
	(0.11)	(0.10)	(0.10)	(0.09)	(0.09)	(0.10)	(0.08)	(0.07)	(0.06)	(0.05)
BIC	3.76	2.86	2.23	1.86	1.68	2.23	2.23	2.22	2.26	2.29
	(0.27)	(0.18)	(0.14)	(0.11)	(0.09)	(0.14)	(0.12)	(0.10)	(0.09)	(0.08)
AICc	3.16	2.64	2.17	1.85	1.68	2.17	2.20	2.19	2.22	2.25
	(0.18)	(0.15)	(0.13)	(0.11)	(0.09)	(0.13)	(0.11)	(0.10)	(0.08)	(0.08)
RIC	3.91	2.94	2.19	1.87	1.73	2.19	2.23	2.22	2.26	2.30
	(0.28)	(0.19)	(0.14)	(0.11)	(0.09)	(0.14)	(0.12)	(0.10)	(0.09)	(0.08)
CAIC	2.10	1.63	1.41	1.31	1.28	1.41	1.38	1.38	1.35	1.32
	(0.11)	(0.08)	(0.07)	(0.05)	(0.04)	(0.07)	(0.05)	(0.04)	(0.03)	(0.03)
CBIC	1.87	1.36	1.16	1.15	1.25	1.16	1.17	1.17	1.15	1.14
	(0.13)	(0.06)	(0.04)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)	(0.02)

of model averaging using the data perturbation approach are not sensitive to the choice of τ under the frameworks of geostatistical models. Therefore, a nonadaptive $\tau = 0.1$ throughout the simulation scenario I was acceptable. Shen and Huang (2006) further developed a methodology about the adaptive choice of τ or the optimal choice of τ under different data sets, although the computation is more time-consuming.

5.3. Simulation scenario II

To evaluate the performance of $\widehat{\mathrm{GIM}}(\lambda)$ in (4.1) for geostatistical regression model selection more realistically, the second simulation scenario was designed based on a data example.

We considered a weather data set in the state of Colorado (Reich and Davis (2008); Chu, Zhu and Wang (2011)), where the response variable is the March mean precipitation (inches per 24-hour period) on the log scale, from 261 weather stations. For each weather station, $p = 10$ covariates are available. In addition to elevation (x_1), slope (x_2), and aspect (x_3), seven spectral bands from a MODIS satellite imagery, B1M–B7M, are available and denoted by x_4, x_5, x_6, x_7, x_8, x_9, and x_{10}. Under (2.2) with the covariance matrix of $\eta(\cdot)$ defined in (5.1), we considered the underlying true models $\beta_0 + \sum_{j=1}^{k} \beta_j x_j(\boldsymbol{s})$ for $k = 1, 5$, and 9.

Table 4. Regression coefficients of three true models (k) of simulation scenario II based on the precipitation data set.

	Regression coefficients									
k	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_9	β_9
1	1.2950	0.5716	0	0	0	0	0	0	0	0
5	1.2948	0.4554	0.0344	−0.0027	−0.1395	0.0139	0	0	0	0
9	1.3105	0.4425	0.0317	−0.0085	0.1902	−0.0108	−0.0265	−0.2487	0.1269	−0.1769

Table 5. Parameter values in the covariance matrix of the data vector and the SNR values for three true models (k) of simulation scenario II based on the precipitation data set.

	Parameter values				
k	a	ν	σ^2_η	σ^2_ε	SNR
1	0.0985	0.5	0.2153	0.2119	2.56
5	0.1220	0.5	0.1497	0.2281	1.66
9	0.1435	0.5	0.1745	0.2281	2.27

For each k, the regression coefficients, $\boldsymbol{\theta} = (\sigma^2_\eta, a, \nu)'$, and σ^2_ε were estimated by REML based on the precipitation data set. These estimated parameter values are displayed in Tables 4 and 5, and were used as the true parameter values of (2.2) when simulating new responses. In simulation scenario II, the spatial random error process $\eta(\cdot)$ was a zero-mean Gaussian stationary process, with covariance matrix based on the exponential covariance function, $\nu = 0.5$ in (5.1), with the range parameter a being around $0.1 \sim 0.14$ as shown in Table 5. The left side of Figure 1 shows a realization of $\eta(\cdot)$ with $(\sigma^2_\eta, a, \nu) = (1, 0.1, 0.5)$. Compared with scenario I, scenario II corresponds to a weaker spatial dependence. In addition, the SNR values for $k = 1$, 5, and 9 are shown in Table 5. Compared with the SNR = 4 in scenario I, the SNR value in scenario II is smaller, indicating that the noise is stronger.

In each simulated data set, a sample size of $n = 100$ was drawn from 261 weather stations based on a simple random sampling scheme. For each of $k = 1$, 5, and 9, we applied six model selection criteria, AIC, BIC, AICc, RIC, CAIC, and CBIC, to select models among all the possible combinations of covariates. For CAIC and CBIC, special cases of the CGIC in (3.4), an estimate $\hat{\sigma}^2_\varepsilon$ of σ^2_ε was obtained by REML based on the full model, and so is invariant to the model choice. The estimate $\hat{\sigma}^2_\varepsilon$ was used in the data perturbation procedure when computing $\widehat{\mathrm{GIM}}(\lambda)$ of (4.1). Throughout scenario II, the model parameters were estimated by REML.

Table 6. Average values of generalized instability measure (GIM) of six information criteria (AIC, BIC, AICc, RIC, CAIC, CBIC) under three true models (k) (left panel) for simulation scenario II and sensitivity analysis of the proposed generalized instability measure (GIM) with respect to various perturbation sizes τ for $k = 5$ (right panel) based on 200 simulation replicates and the precipitation data set under simulation scenario II. The values in parentheses are the corresponding standard errors.

Criterion	k			τ		
	1	5	9	0.1	0.5	0.9
AIC	4.12	3.88	3.18	3.88	3.64	3.54
	(0.39)	(0.35)	(0.27)	(0.35)	(0.24)	(0.17)
BIC	4.85	4.67	3.54	4.67	4.16	4.00
	(0.50)	(0.45)	(0.30)	(0.45)	(0.32)	(0.23)
AICc	4.19	3.98	3.24	3.98	3.73	3.64
	(0.40)	(0.36)	(0.27)	(0.36)	(0.25)	(0.18)
RIC	4.81	4.63	3.48	4.63	4.07	3.90
	(0.49)	(0.44)	(0.30)	(0.44)	(0.31)	(0.21)
CAIC	2.49	2.22	1.56	2.22	2.11	1.98
	(0.23)	(0.21)	(0.16)	(0.21)	(0.14)	(0.10)
CBIC	1.41	1.59	1.24	1.59	1.64	1.63
	(0.16)	(0.18)	(0.14)	(0.18)	(0.13)	(0.09)

5.4. Simulation result II

To assess the instability of model selection based on AIC, BIC, AICc, RIC, CAIC, and CBIC, the corresponding values of $\widehat{\mathrm{GIM}}(\lambda)$ in (4.1) were computed based on the data perturbation illustrated in Section 4.1; the perturbation size was set to $\tau = 0.1$ for computing $\widehat{\mathrm{ISI}}(\lambda)$. The left panel of Table 6 shows the results of $\widehat{\mathrm{GIM}}(\lambda)$ under the AIC, BIC, AICc, RIC, CAIC, and CBIC criteria for $k = 1$, 5, and 9 based on 200 replicates. The conditional information criteria, CAIC and CBIC, tend to have smaller GIM values than unconditional information criteria in this simulation. This indicates that the conditional information criteria are more stable, and thus are more suitable, than unconditional information criteria in the geostatistical regression model selection. We again conclude that model selection and model averaging give more comparable results as the uncertainty of model selection decreases.

We performed a sensitivity analysis to evaluate the effect of the perturbation size τ on the computation of $\widehat{\mathrm{GIM}}(\lambda)$. The average values of $\widehat{\mathrm{GIM}}(\lambda)$ and the corresponding standard errors based on $k = 5$ and 200 simulation replicates are given in the right panel of Table 6. For a given model selection criterion, the numerical results indicate that the computation of $\widehat{\mathrm{GIM}}(\lambda)$ is not sensitive to the different choices of τ. An approximately unbiased estimator $\widehat{\mathrm{ISI}}(\lambda)$ of $\mathrm{ISI}(\lambda)$

can be obtained based on a data perturbation technique as $\tau \to 0^+$ (Huang and Chen (2007)), but it may produce numerical instability if τ is too small (e.g., $\tau = 10^{-10}$). Therefore, we used $\tau = 0.1$ throughout the simulation scenarios I and II, resulting in a somewhat larger bias but a smaller variance (e.g., Shen and Huang (2006); Huang and Chen (2007); Chen, Yang and Li (2014)).

The results from simulation scenarios I and II suggest that the size of the underlying true model impacts the instability of geostatistical regression model selection. In practice, we can combine the conditional information criteria (e.g., CAIC and CBIC), the RIC criterion, and the empirical rule in Section 4.2 to jointly infer the size of the underlying true model. If the number of covariates in the underlying true model is relatively small, a model selection procedure is relatively unstable and a model averaging method is preferred for more accurate spatial prediction, while the conditional information criterion or the RIC criterion is recommended for variable selection. When selection uncertainty decreases (see, e.g., the left panel of Table 3 and Figure 2(b)), model selection and model averaging give comparable prediction results (see, e.g., Table 1 and Figure 2(a)). Thus, prediction and variable selection might be based on a model selection criterion, where the conditional information criteria are still preferable.

6. Precipitation Data Example

We applied our proposed methodology to the precipitation data set in the state of Colorado (Reich and Davis (2008); Chu, Zhu and Wang (2011)) as illustrated in Section 5.3. We considered all possible combinations of covariates to investigate the size of the underlying geostatistical regression model.

Guided by the simulation study, we applied RIC, CAIC, CBIC, and the empirical rule developed in Section 4.2 to jointly estimate the size of the underlying geostatistical regression model. Table 7 provides the three best models selected by RIC, CAIC, and CBIC among 2^{10} candidate models. The estimated results of $\text{ReGIM}_j(\lambda)$ of (4.3) for $j = 1, \ldots, 10$ are shown in Figure 3. For estimating $\text{ReGIM}_j(\lambda)$, an estimate of $\text{GIM}_j(\lambda)$ of (4.2) was computed based on the estimation procedure of $\text{GIM}(\lambda)$ in Section 4.1 with $\mathcal{M}$ replaced by $\mathcal{P}_j$; $j = 1, \ldots, 10$. The results in Table 7 indicate that there should be 2 to 3 important covariates in the geostatistical regression model for precipitation. Two covariates (elevation and B4M) were selected by RIC, which agrees with a previous analysis (Chu, Zhu and Wang (2011)), whereas three covariates (elevation, slope and B1M) were selected by CAIC and two covariates (slope and B7M) were selected by CBIC.

Table 7. Selected models with the corresponding covariates ("Yes") among 2^{10} candidate models for RIC, CAIC, and CBIC applied to the precipitation data example.

Criterion	Covariate									
	Elevation	Slope	Aspect	B1M	B2M	B3M	B4M	B5M	B6M	B7M
RIC	Yes	-	-	-	-	-	Yes	-	-	-
CAIC	Yes	Yes	-	Yes	-	-	-	-	-	-
CBIC	-	Yes	-	-	-	-	-	-	-	Yes

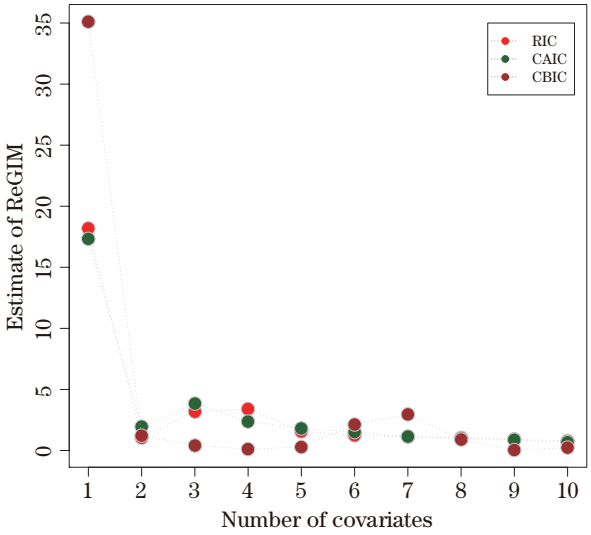

Figure 3. Estimates of relative difference of generalized instability measures (ReGIM) for RIC, CAIC, and CBIC applied to the precipitation data example.

It is clear that $\widehat{\text{ReGIM}}_j(\lambda)$ becomes more stable when j is more than 2, as the elbow pattern in Figure 3 indicates. These results suggest that the size of the underlying geostatistical regression model for this data set is relatively small and thus model averaging would be more suitable for predicting precipitation at unsampled locations. Further, RIC, CAIC, and CBIC are the recommended model selection criteria for choosing important covariates.

7. Conclusions and Discussion

In this paper, a GIM criterion is proposed to measure the instability of geostatistical regression model selection. The proposed criterion takes into account the uncertainties of model selection and parameter estimation and thus more accurately reflects the complexity of a given model fitting procedure.

From the results of two simulation scenarios and a data example, we recom-

mend that the conditional information criteria (e.g., CAIC and CBIC), the RIC criterion, and the empirical rule developed in Section 4.2 can be used together in practice to estimate the size of the underlying geostatistical regression model. When model selection uncertainty is relatively large, model averaging method is preferred when spatial prediction is the main interest, and the conditional information criterion or the RIC criterion is recommended for identifying important covariates. As model selection uncertainty decreases, the predictive performance by model selection and model averaging tends to be similar. Thus, prediction and selection of covariates can both be obtained from a model selection criterion, where the conditional information criterion is preferable.

Although our GIM criterion is developed in the setting of geostatistical regression models, it could also explore model selection or model averaging under such frameworks as linear mixed-effects models. For parameter estimation, it is known that parameters in the Matérn correlation function cannot be estimated well, even when increasing amounts of data are collected densely in a fixed domain. Irvine, Gitelman and Hoeting (2007) have compared the performance of ML and REML estimates of spatial covariance parameters under various sampling designs via simulation studies and showed that these estimates still have room for improvement. While our focus is on the selection of covariates, the instability of the estimation of spatial covariance parameters is an important issue to address and could be explored based on our GIM criterion.

A limitation of the proposed method is its computational intensity when dealing with massive data sets or a large number of covariates, because the estimation procedure of GIM(λ) in Section 4.1 involves inverting an $n \times n$ covariance matrix and selecting a best model from a large class of candidate models. Computationally more efficient methods can be considered such as least angle regression (e.g., Efron et al. (2004)), covariance tapering (e.g., Furrer, Genton and Nychka (2006)), and low-rank methods (e.g., Cressie and Johannesson (2008)).

Acknowledgment

We thank the Editor, an associate editor, and two referees for their helpful comments and suggestions. We would like to thank Professor Jennifer A. Hoeting for her detailed comments, which have been very helpful in improving the article. The research of Chun-Shu Chen was supported by the Ministry of Science and Technology of Taiwan under Grant MOST 103-2118-M-018-002, the research of Jun Zhu was supported by US Department of Interior USGS CESU Award

G16AC00344, and the research of Tingjin Chu was supported by the National Natural Science Foundation of China (grant 11301536).

Appendix

Proof of Proposition 1. Since $\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) = \boldsymbol{X}_M(\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}$, we rewrite $\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M)$ of (2.9) as

$$\begin{aligned}
\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M) &= \hat{\sigma}_{\varepsilon,M}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M(\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1} \\
&= \hat{\sigma}_{\varepsilon,M}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1} \\
&= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) \\
&\quad +\hat{\sigma}_{\varepsilon,M}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) \\
&= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big(\boldsymbol{I} - \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\big) + (\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M) \\
&\quad +\hat{\sigma}_{\varepsilon,M}^2\boldsymbol{I})(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M) + \hat{\sigma}_{\varepsilon,M}^2\boldsymbol{I})^{-1}\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M) \\
&= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big(\boldsymbol{I} - \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\big) + \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\ .
\end{aligned}$$

It follows that

$$\begin{aligned}
\operatorname{tr}\Big(\hat{\boldsymbol{H}}_M(\hat{\boldsymbol{\theta}}_M)\Big) &= \operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big(\boldsymbol{I} - \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\big)\Big) + \operatorname{tr}\Big(\hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\Big) \\
&= \operatorname{tr}\Big(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_M)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big(\boldsymbol{I} - \hat{\boldsymbol{Q}}_M(\hat{\boldsymbol{\theta}}_M)\big)\Big) + |M|\ .
\end{aligned}$$

Combining this with (3.4), we obtain the desired result.

Proof of Proposition 2. Following (3.3), (3.6), and the definition of ISI(λ) in (3.12), we have

$$\begin{aligned}
&\delta^{-1}\Big\{\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i) - \hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\Big\} \\
&= \delta^{-1}\Big\{\Big[\hat{\boldsymbol{H}}_{\hat{\gamma}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})(\boldsymbol{Z}+\delta\boldsymbol{e}_i)\Big]_i - \Big[\hat{\boldsymbol{H}}_{\hat{\gamma}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})\boldsymbol{Z}\Big]_i\Big\} = \Big[\hat{\boldsymbol{H}}_{\hat{\gamma}(\lambda)}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})\Big]_{ii} \\
&= \hat{\sigma}_{\varepsilon,\hat{\gamma}(\lambda)}^2\left[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}\left\{\boldsymbol{X}_{\hat{\gamma}(\lambda)}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}\right\}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\right]_{ii} + \Big[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\Big]_{ii},
\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)}) + \hat{\sigma}_{\varepsilon,\hat{\gamma}(\lambda)}^2\boldsymbol{I}$, $[\boldsymbol{A}]_i$ denotes the ith element of vector $\boldsymbol{A}$, and $[\boldsymbol{B}]_{ii}$ denotes the ith diagonal element of matrix $\boldsymbol{B}$. Let $\boldsymbol{P}$ and $\boldsymbol{G}$ be $n\times n$ positive definite matrices and let $\boldsymbol{R}$ be an $n\times m$ matrix with $\operatorname{rank}(\boldsymbol{R}) = m$. Basic properties of positive definite matrices (e.g., Harville (1997)) are used in proving Proposition 2: (i) $\boldsymbol{P}^{-1}$ and $\boldsymbol{G}^{-1}$ are positive definite; (ii) $\boldsymbol{R}'\boldsymbol{P}\boldsymbol{R}$ and $\boldsymbol{R}'\boldsymbol{G}\boldsymbol{R}$ are positive definite; (iii) $\boldsymbol{P}\boldsymbol{G}\boldsymbol{P}$ and $\boldsymbol{G}\boldsymbol{P}\boldsymbol{G}$ are positive definite; (iv) If $\boldsymbol{P}\boldsymbol{G} = \boldsymbol{G}\boldsymbol{P}$, then $\boldsymbol{P}\boldsymbol{G}$ is also positive definite. Because $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}$ are $n\times n$ positive definite matrices and $\boldsymbol{X}_{\hat{\gamma}(\lambda)}$ is an $n\times(|\hat{\gamma}(\lambda)|+1)$ design matrix with $\operatorname{rank}(\boldsymbol{X}_{\hat{\gamma}(\lambda)}) = |\hat{\gamma}(\lambda)|+1$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}(\boldsymbol{X}_{\hat{\gamma}(\lambda)}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)})^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}'\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}$ is

positive definite by (i)-(iii), and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}$ is positive definite by (i) and (iv). As a consequence, $\big[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)}(\boldsymbol{X}'_{\hat{\gamma}(\lambda)}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\boldsymbol{X}_{\hat{\gamma}(\lambda)})^{-1}\boldsymbol{X}'_{\hat{\gamma}(\lambda)}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big]_{ii} > 0$ and $\big[\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}^{-1}\big]_{ii} > 0$ for $i = 1, \ldots, n$. Thus,

$$
\begin{aligned}
\mathrm{ISI}(\lambda) &= E\left[\lim_{\delta\to 0}\delta^{-1}\sum_{i=1}^{n}\left|\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i)-\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\right|\right]\\
&= \sum_{i=1}^{n}\lim_{\delta\to 0}\delta^{-1}E\left[\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i)-\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\right]\\
&= \sum_{i=1}^{n}\lim_{\delta\to 0}\delta^{-1}\left[E E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i)\big|\boldsymbol{S}\big)-E E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\big|\boldsymbol{S}\big)\right]\\
&= \sum_{i=1}^{n}E\left[\lim_{\delta\to 0}\delta^{-1}\left(E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z}+\delta\boldsymbol{e}_i)\big|\boldsymbol{S}\big)-E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\big|\boldsymbol{S}\big)\right)\right]\\
&= \sum_{i=1}^{n}E\left(\frac{\partial}{\partial S(\boldsymbol{s}_i)}E\big(\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})\big|\boldsymbol{S}\big)\right),
\end{aligned}
$$

where $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\boldsymbol{Z})$ and $\hat{S}_{\hat{\gamma}(\lambda)}(\boldsymbol{s}_i;\hat{\boldsymbol{\theta}}_{\hat{\gamma}(\lambda)})$ are the same as in (3.12). This completes the proof.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory* (V. Petrov and F. Csáki eds.), Akademiai Kiádo, Budapest, 267-281.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Stat.* **24**, 2350-2383.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (second edition). Springer-Verlag, New York.

Chen, C.-S. and Huang, H.-C. (2012). Geostatistical model averaging based on conditional information criteria. *Environ. Ecol. Stat.* **19**, 23-35.

Chen, C.-S., Yang, H.-D. and Li, Y. (2014). A stabilized and versatile spatial prediction method for geostatistical models. *Environmetrics* **25**, 127-141.

Chu, T., Zhu, J. and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *Ann. Stat.* **39**, 2607-2625.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Stat. Sci.* **19**, 81-94.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **70**, 209-226.

Cressie, N. and Lahiri, S. N. (1993). The asymptotic distribution of REML estimators. *J. Multivar. Anal.* **45**, 217-233.

Cressie, N. and Lahiri, S. N. (1996). Asymptotics for REML estimation of spatial covariance parameters. *J. Stat. Plan. Infer.* **50**, 327-341.

Efron, B. (2014). Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* **109**, 991-1007.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Stat.* **32**, 407-499.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Stat.* **22**, 1947-1975.

Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *J. Comput. Graph. Stat.* **15**, 502-523.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881-889.

Ghosh, D. and Yuan, Z. (2009). An improved model averaging scheme for logistic regression. *J. Multivar. Anal.* **100**, 1670-1681.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer, New York.

Hoeting, J. A., Davis, R. A., Merton, A. A. and Thompson, S. E. (2006). Model selection for geostatistical models. *Ecol. Appl.* **16**, 87-98.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Stat. Sci.* **14**, 382-401.

Huang, H.-C. and Chen, C.-S. (2007). Optimal geostatistical model selection. *J. Am. Stat. Assoc.* **102**, 1009-1024.

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

Irvine, K. M., Gitelman, A. I. and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *J. Agric. Biol. Environ. Stat.* **12**, 450-469.

Johnson, D. S. and Hoeting, J. A. (2011). Bayesian multimodel inference for spatial regression models. *PLoS ONE* **6(11)**: e25677.

Matérn, B. (2013). *Spatial Variation*. Springer Science & Business Media.

McGilchrist, C. A. (1989). Bias of ML and REML estimators in regression models with ARMA errors. *J. Stat. Comput. Simul.* **32**, 127-136.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**, 179-191.

Reich, R. and Davis, R. (2008). *Lecture Notes of Quantitative Spatial Analysis*. Colorado State University, Fort Collins, CO.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.

Shao, J. (1997). An asymptotic theory for model selection (with discussion). *Stat. Sinica* **7**, 221-264.

Shen, X. and Huang, H.-C. (2006). Optimal model assessment, selection and combination. *J. Am. Stat. Assoc.* **101**, 554-568.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *J.*

Am. Stat. Assoc. **93**, 120-131.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models: When and how? *J. Am. Stat. Assoc.* **100**, 1202-1214.

Institute of Statistics and Information Science, National Changhua University of Education, Changhua 500, Taiwan.

E-mail: cschen@cc.ncue.edu.tw

Department of Statistics and Department of Entomology, University of Wisconsin, Madison, WI 53706, USA.

E-mail: jzhu@stat.wisc.edu

Center for Applied Statistics and Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: tingjin_chu@outlook.com

(Received August 2016; accepted December 2016)