NIASRA
NATIONAL INSTITUTE FOR APPLIED
STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong

Working Paper

03-14

Evaluation of Diagnostics for Hierarchical Spatial Statistical Models

Noel Cressie and Sandy Burden

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email:
anica@uow.edu.au

1

Evaluation of Diagnostics for Hierarchical Spatial Statistical Models

Noel Cressie and Sandy Burden

National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong

1.1 Introduction

In the twenty-first century, we are able to build large, complex statistical models that are very much like the scientific processes they represent. We use diagnostics to highlight inadequacies in the statistical model, and because of the complexity many different diagnostics are needed. This is analogous to the process of diagnosis in the medical field, where a suite of diagnostics is used to assess the health of a patient.

This chapter is focussed on *evaluating* model diagnostics. In the medical literature, a structured approach to diagnostic evaluation is used, based on measurable outcomes such as Sensitivity, Specificity, ROC curves, and False Discovery Rate. We suggest using the same framework to evaluate model diagnostics for hierarchical spatial statistical models; we note that the concepts are the same in the non-spatial and non-hierarchical setting, although the specific proposals given in this chapter may be difficult to generalize.

1.1.1 Hierarchical spatial statistical models

The statistical models that we use to model a spatial process involve many sources of uncertainty, including uncertainty due to the observation process, uncertainty in the spatial process, and uncertainty in the parameters. A hierarchical spatial model allows us to express these uncertainties in terms of conditional probabilities that define respectively, the data model, the process model, and the parameter model (e.g., Cressie and Wikle 2011, Chapter 2).

This is a Book Title Name of the Author/Editor

Suppose that $Y \equiv \{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a spatial process of scientific interest, where $\mathcal{D}$ is a known region in the d-dimensional Euclidean space $\mathbb{R}^d$. We use a spatial statistical model that depends on unknown parameters, $\boldsymbol{\theta}_p$, to quantify our uncertainty in the scientific process of interest, and we use a data model that depends on unknown parameters, $\boldsymbol{\theta}_d$, to quantify our uncertainty in the measurement process. The joint distribution of Y, given all possible parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\theta}_d^T, \boldsymbol{\theta}_p^T)^T$, can be written as,

$$[Y|\boldsymbol{\theta}] = [Y|\boldsymbol{\theta}_p], \tag{1.1}$$

where $[A|B]$ is generic notation for the probability density or mass function of A given B. We call (1.1) the *process model.*

Due to measurement error and incomplete sampling, the scientific process is not directly observed. Instead, $\boldsymbol{Z} \equiv (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))^T$ is observed, whose uncertainty can be quantified through the *data model*,

$$[\boldsymbol{Z}|Y, \boldsymbol{\theta}] = [\boldsymbol{Z}|Y, \boldsymbol{\theta}_d]. \tag{1.2}$$

In a fully Bayesian model, uncertainty in the parameters is quantified through a parameter model,

$$[\boldsymbol{\theta}] = [\boldsymbol{\theta}_d, \boldsymbol{\theta}_p], \tag{1.3}$$

where recall that $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_d$ are the parameters from the process model and the data model, respectively.

The use of conditional distributions to specify a hierarchical statistical model is a powerful way to model complex dependence structures with many sources of uncertainty. Using Bayes' Rule, the posterior distribution for the process and the parameters, which forms the basis for inference in a Bayesian hierarchical model, is given by,

$$[Y, \boldsymbol{\theta}|\boldsymbol{Z}] = [\boldsymbol{Z}|Y, \boldsymbol{\theta}][Y|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Z}]. \tag{1.4}$$

Statistical modelling is commonly undertaken to make inference on (i.e., predictions for) the spatial process Y. The usefulness of the hierarchical framework is demonstrated by comparison with a non-hierarchical-model specification. Bayesian, non-hierarchical statistical models implicitly integrate over the process model to obtain the posterior distribution, $[\boldsymbol{\theta}|\boldsymbol{Z}] = \int_Y [\boldsymbol{Z}|Y, \boldsymbol{\theta}][Y|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Z}]dY$. When Y is not included in the model specification, the scientific relationships and the observation process are confounded. This has important implications for diagnostics because uncertainty in the measurement process is very different from uncertainty in the scientific process.

1.1.2 Diagnostics

Once we have specified a hierarchical spatial statistical model and fitted it to the data $\boldsymbol{Z}$, we use diagnostics to "stress-test" the model, to assess whether it is adequate for our purposes. There are a wide range of diagnostics that we may use to do this, because the meaning of "adequate" depends on the purpose of fitting the model in the first place. Analogous to a medical diagnostic, each model diagnostic should be looking for something unusual, to indicate an inadequacy in the model.

The general features of common statistical-model diagnostics are well known and found in many statistical texts (e.g., Carlin and Louis 2009; Gelman et al. 2013; Huber-Carol et al.

2002), including those for hierarchical models (e.g., Banerjee et al. 2004; Cressie and Wikle 2011) and those for spatial data (e.g., Cressie 1993; Gelfand et al. 2010; Schabenberger and Gotway 2005). They include diagnostics to assess residuals (e.g., Belsley et al. 1980; Cook and Weisberg 1982; Cox and Snell 1968; Fox 1991; Kaiser et al. 2012), parameter estimates (e.g., Bousquet 2008; Evans and Moshonov 2006; Presanis et al. 2013), modelling assumptions (e.g., Goel and De Groot 1981; O'Hagan 2003; Scheel et al. 2011), and prior distributions (e.g., Hill and Spall 1994).

Many diagnostic criteria derive from probability measures (e.g., Crespi and Boscardin 2009; Meng 1994; Steinbakk and Storvik 2009), which may or may not be associated with an explicit hypothesis test. Alternatives include visualising a diagnostic (e.g., Bradley and Haslett 1992; Massmann et al. 2014; Murray et al. 2013) and identifying "interesting" values heuristically or using an empirically derived "rule of thumb."

For hierarchical models, we typically wish to diagnose the adequacy of the model fitted to $[Y|\boldsymbol{\theta}_p]$. However, Y is not observed. Instead we observe data $\boldsymbol{Z}$, which includes measurement error and possible summarisation and approximation. Loy and Hofmann (2013), Yan and Sedransk (2007), and Yuan and Johnson (2012) are general references, and an important class of hierarchical-model diagnostics is based on predictive distributions (e.g., Box 1980; Gelfand et al. 1992; Gelman et al. 1996; O'Hagan 2003).

Diagnostics for spatial statistical models (e.g., Anselin and Rey 2010; Christensen et al. 1992; Cressie 1993; Cressie and Wikle 2011; Gelfand et al. 2010; Glatzer and Müller 2004) are more complex due to spatial dependence between locations (e.g., Baddeley et al. 2005; Kaiser et al. 2012; Lee and Ghosh 2009). Global diagnostics applied to the fitted model give an indication of the overall adequacy of the model, but they do not assess the fit of the model at particular locations (e.g., Hering and Genton 2011). Here, local statistics can be powerful diagnostics (see Fotheringham 2009; Fotheringham and Brunsdon 1999, for a review of local analysis), although they can be computationally expensive. Examples include the local indicators of spatial association (LISA) (Anselin 1995; Getis and Ord 1992; Moraga and Montes 2011; Ord and Getis 1995), LICD, a LISA equivalent for categorical data (Boots 2003), the structural similarity index (SSM) (Robertson et al. 2014; Wang et al. 2004), the S-statistic (Karlström and Ceccato 2002), the local spatial heteroskedasticity statistic (LOSH) (Ord and Getis 2012; Xu et al. 2014) and local diagnostics based on the spatial scan statistic for identifying clusters (Kulldorff et al. 2006; Read et al. 2013).

1.1.3 Evaluation

Model diagnostics are widely used, and questions such as: "How reliable are the results of the diagnostic?" and "What are the consequences of using a fitted model that a particular diagnostic deemed inadequate?" naturally arise. In the statistical literature, these questions are answered in ways that include reference to theoretical properties of the diagnostic (e.g., Gneiting 2011; Robins et al. 2000), the performance of the diagnostic on simulated data with known properties (e.g., Dormann et al. 2007), and the distribution of p-values (e.g., He et al. 2013). When a diagnostic is evaluated using the same data that were used to fit the model, the results are well known to be biased (Bayarri and Berger 2000; Dahl 2006; Efron 1986; Hjort et al. 2006). An alternative is to use cross-validation (Gelfand 1996; Le Rest et al. 2014; Stone 1974; Zhang and Wang 2010), where the model is fitted to

$m < n$ observations and evaluated using the remaining $n - m$ observations. While cross-validation is considered a gold standard for diagnostics (Gelfand et al. 1992; Marshall and Spiegelhalter 2003; Stern and Cressie 2000), it is computationally expensive and may be impractical for very large datasets. Alternatives such as testing datasets (Efron 1983, 1986), importance sampling (Stern and Cressie 2000), simulation-based model checking (Dey et al. 1998), posterior predictive checks (Gelman et al. 1996; Marshall and Spiegelhalter 2007), and approaches that balance bias with the computational burden of cross-validation (Bayarri and Berger 2000; Bayarri and Castellanos 2007) may also be used.

For hierarchical spatial statistical models, an obvious class of diagnostics identifies those locations where the model is inadequate and those locations where it is adequate. However, in most cases the diagnostic will misclassify some locations. There is potentially a strong parallel here between spatial-model diagnostics and medical diagnostics (e.g., Moraga and Montes 2011; van Smeden et al. 2014), where a diagnostic test is used to identify unusual values (e.g., Pepe and Thompson 2000; Sackett and Haynes 2002). Two summary statistics that are routinely used to assess the performance of medical diagnostics are Sensitivity and Specificity (e.g., Akobeng 2007; Enøe et al. 2000; Hui and Zhou 1998). More recently, there has been a greater use of the False Discovery Rate (FDR) (e.g., Benjamini and Hochberg 1995, 1997; Efron 2004; Genovese and Wasserman 2002; Storey 2003; Storey and Tibshirani 2003), and the False Nondiscovery Rate (FNR) (e.g., Craiu and Sun 2008). FDR has been used with correlated data (Benjamini and Yekutieli 2001; Finner et al. 2007; Hu et al. 2010) and, for spatial data, generalised degrees of freedom and clustering may be used to increase the power of the FDR approach (Benjamini and Heller 2007; Shen et al. 2002).

In Section 1.2, we introduce a simple example of county-level Sudden Infant Death Syndrome (or cot death) to illustrate our ideas. In Section 1.3, we exploit a strong analogy between medical diagnosis and model diagnosis, and we define the summary measures of Specificity, Sensitivity, False Discovery Rate, and False Nondiscovery Rate for evaluating a diagnostic. In Section 1.4, we use these ideas to define a Discovery curve that can be interpreted in an analogous way to the Receiver Operating Characteristic (ROC) curve. Finally, a discussion and our conclusions are given in Section 1.5.

1.2 Example: Sudden Infant Death Syndrome (SIDS) Data for North Carolina

This section introduces an example that will be used to illustrate our proposal for the evaluation of model diagnostics. The dataset includes the counts of Sudden Infant Death Syndrome (SIDS) for the 100 counties of North Carolina for the period July 1, 1974 – June 30, 1978 (Cressie 1993; Cressie and Chan 1989; Symons et al. 1983), where the counties are numbered according to the alphabetical order of their county name. For each county, the dataset also includes the number of live births, the spatial location of the county (here specified as the county centroid), and the adjacent counties (i.e., all pairs whose county seats are within 30 miles of each other); see Figure 1.1. The SIDS data have been extensively studied (e.g., Bivand 2014; Cressie 1993; Cressie and Chan 1989; Cressie and Read 1985; Sengupta and Cressie 2013), and they are widely available (e.g., in the spdep package in the R Statistical Software, Bivand 2014; R Core Team 2014).

Our purpose in this chapter is not to identify new diagnostics nor in this section to model the SIDS data in a new way. Instead, we shall model the data with a simple statistical

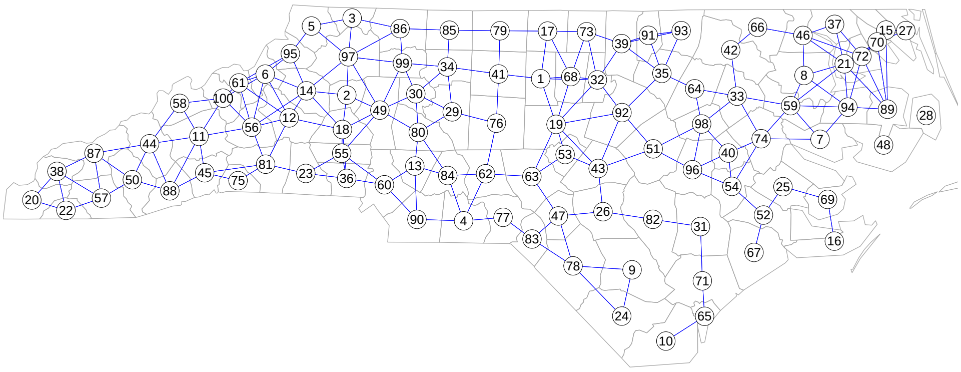

Figure 1.1 Map of the 100 counties in North Carolina, showing edges between counties whose seats are within 30 miles of each other. The counties are numbered according to the alphabetical order of their county name. *Figure adapted from Bivand (2014).*

model and then diagnose the fit of the model using several established diagnostics. Using these results, we shall then evaluate the diagnostics for the model in the manner described in Sections 1.3 and 1.4. For this reason, we base our analysis on the results of previous exploratory analyses conducted by Cressie and Read (1985), Cressie and Chan (1989), and Cressie (1993, Sections 4.4, 6.2, and 7.6). These authors found that the Freeman-Tukey (square-root) transformation of the SIDS rates stabilises the variance and results in a symmetrical distribution, so that an approximate Gaussian assumption can be made for the transformed data. Most analyses of this transformed dataset are based on an auto Gaussian spatial model. We will follow this approach and fit a null statistical model that assumes a constant mean and Gaussian variation in the error. All 100 counties are included; note that in the past, Anson County (county 4) has been identified as an outlier and sometimes removed. Having fitted the model, we use the local Moran I statistic and the local Getis-Ord G^* statistic to assess the adequacy of the fitted model. The local statistics will be applied to the residuals to identify whether there is unusual spatial behaviour after the model has been fitted.

In our study, recall that the seat of county i is used to define its location $\boldsymbol{s}_i; i = 1, ..., 100$. Previous studies have found that the spatial correlation between counties is close to zero at distances, $d_{ij} \equiv \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, of 30 miles or more.

For $i = 1, ..., 100$, let $N(\boldsymbol{s}_i)$ and $S(\boldsymbol{s}_i)$ denote the number of live births and the number of SIDS deaths, respectively, for county i. Its Freeman-Tukey transformed SIDS rate (per thousand live births) is given by,

$$Z(\boldsymbol{s}_i) \equiv (1000S(\boldsymbol{s}_i)/N(\boldsymbol{s}_i))^{1/2} + (1000(S(\boldsymbol{s}_i) + 1)/N(\boldsymbol{s}_i))^{1/2} .$$

The null model for the transformed SIDS rate is defined as,

$$Z(\boldsymbol{s}_i) = \mu_0 + \delta(\boldsymbol{s}_i), \tag{1.5}$$

where the mean transformed rate, μ_0, is assumed to be constant, and the error, $\delta(\boldsymbol{s}_i)$, is assumed to have a Gaussian distribution with mean zero and variance var$(\delta(\boldsymbol{s}_i)) = \sigma_\delta^2 V_\delta(\boldsymbol{s}_i)$, for $\sigma_\delta^2 > 0$ and $V_\delta(\boldsymbol{s}_i) \equiv N(\boldsymbol{s}_i)^{-1}$. We fitted this model using weighted least squares, but not generalized least squares since initially $\delta(\cdot)$ is assumed to exhibit no spatial dependence. The estimate for the mean was 2.84 with a standard error of 0.075.

We would now like to determine whether there is any spatial clustering in the residuals after fitting the null model. To do this, we applied the local Moran I statistic (Anselin 1995), and the local Getis-Ord G^* statistic (Getis and Ord 1992) to the residuals from the model. For a spatial process $\{x_i : i = 1, ..., n\}$, the local Moran I statistic is given by,

$$I_i = \frac{(x_i - \bar{x}) \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})}{n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{1.6}$$

where w_{ij} is a measure of the spatial dependence between observations i and j. In this example, the spatial-dependence matrix is given by $\boldsymbol{W} \equiv \{w_{ij} : i, j = 1, ..., 100\}$, where $w_{ii} = 0$; and for $i \neq j$, $w_{ij} = 1$ when $d_{ij} \leq 30$ miles, and $w_{ij} = 0$ otherwise.

The local Getis-Ord G^* statistic is given by,

$$G_i^* = \frac{\sum_{j=1}^{n} c_{ij} x_j - \bar{x} \sum_{j=1}^{n} c_{ij}}{\left(n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 (n \sum_{j=1}^{n} c_{ij}^2 - (\sum_{j=1}^{n} c_{ij})^2)/(n-1) \right)^{1/2}}, \tag{1.7}$$

where the spatial-dependence matrix is given by $\boldsymbol{C} \equiv \{c_{ij} : i, j = 1, ..., 100\}$. In this example, $c_{ii} = 1$; and for $i \neq j$, $c_{ij} = 1$ when $d_{ij} \leq 30$ miles, and $c_{ij} = 0$ otherwise.

Values for the local Moran I statistic and the local Getis-Ord G^* statistic are shown in Figures 1.2 and 1.3; in each case, values of the statistic that are "statistically significant for $\alpha = 0.05$" are highlighted. The local Moran I statistic identifies 18 counties with significant spatial dependence. The local Getis-Ord G^* statistic identifies 12 counties with significant spatial dependence. Using the local Moran I statistic, we would conclude that our model is inadequate for four clusterings of counties in the study area. Using the local Getis-Ord G^* statistic, we would conclude that our model is inadequate for three clusterings of counties in the study area. Both diagnostics identify two common spatial clusterings, but each also identifies additional spatial clusterings of counties.

1.3 Diagnostics as Instruments of Discovery

Whether diagnostics are applied to a spatial model, a hierarchical model, or really any statistical model, they are meant to highlight inadequacies (and adequacies) of the model. While one diagnostic might indicate no inadequacies with a model, it is perfectly plausible that another diagnostic might reveal inadequacies. And just because an inadequacy is found, it does not mean that it is truly an inadequacy. This latter statement may look different from the usual discussion about diagnostics, and it is something we shall pursue in this chapter.

We deem the declaration of an inadequacy of the model a "positive." Likewise the declaration of an adequacy is deemed a "negative." This is clearest in the spatial setting where

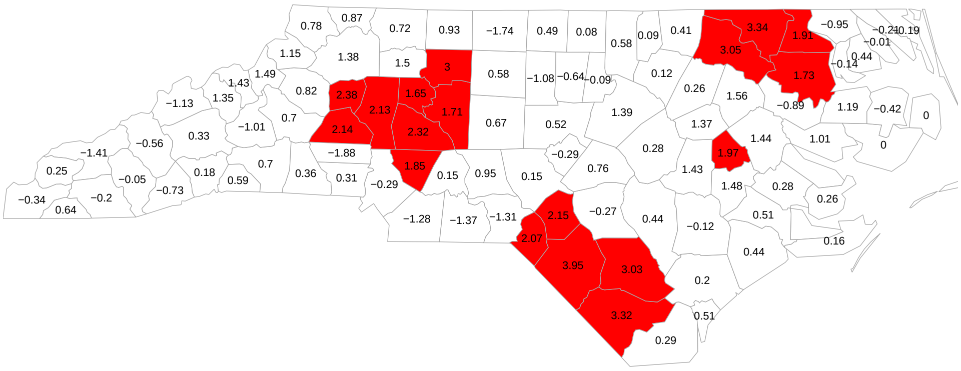

Figure 1.2 Local Moran I statistic for the residuals of the null model fitted using the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

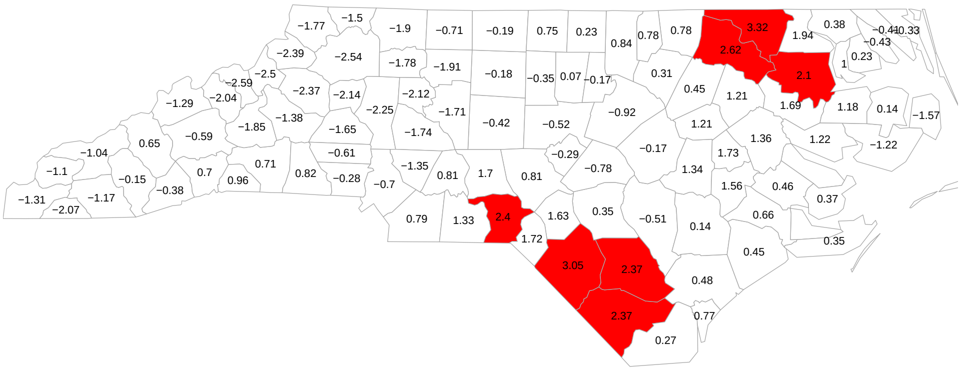

Figure 1.3 Local Getis-Ord G^* statistic for the residuals of the null model fitted using the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

Table 1.1 A 2×2 table resulting from our diagnostic evaluation based on a precise, follow-up reanalysis.

	Negative	Positive	Total
Diagnostic Negative	A_{TN}	A_{FN}	A_N
Diagnostic Positive	A_{FP}	A_{TP}	A_P
Total	$A_{TN}+A_{FP}$	$A_{FN}+A_{TP}$	n

each datum $Z(\boldsymbol{s}_i)$ at spatial location $\boldsymbol{s}_i$, for $i = 1, ..., n$, is potentially a positive (model gives an inadequate fit) or a negative (model gives an adequate fit). If one thinks of diagnosing a model as an act of discovery, analogous to diagnosing a patient in a medical setting (see Section 1.1), then an indication by a diagnostic that something is unusual is seen as a positive.

Discovery of positives and negatives comes with its own uncertainty; a negative could either be a "true negative (TN)" or a "false negative (FN)," and a positive could either be a "false positive (FP)" or a "true positive (TP)." In the spatial setting, if we have n data points and we diagnose the adequacy of each one, then the number of positives (A_P) plus the number of negatives (A_N) equals n. From the discussion above, we have

$$\begin{aligned} A_{TN} + A_{FN} &= A_N \\ A_{FP} + A_{TP} &= A_P, \end{aligned} \tag{1.8}$$

where $A_N + A_P = n$, and clearly A_{TN} is the number of *True negatives*, A_{FN} is the number of *False negatives*, A_{FP} is the number of *False positives*, and A_{TP} is the number of *True positives*.

The way equation (1.8) is written suggests Table 1.1, which is a 2×2 table where the rows are classified according to the behaviour of the diagnostic; negatives along the first row and positives along the second row. The columns are classified according to a precise, "follow-up" reanalysis of each spatial datum; down the first column are the follow-up negatives and down the second column are the follow-up positives. Hence, the top left-hand corner gives the number of True negatives (since both row and column correspond to negatives); the top right-hand corner gives the number of False negatives (since the row is negative but the column shows it should actually be positive); and so forth.

This chapter is about *evaluating* diagnostics and is not directly concerned with defining a "better" diagnostic. Although, once we have a yard-stick by which to compare diagnostics, there is a path forward to making them better and better. Our strategy is to take a given diagnostic, based on a particular fitted spatial model, and to determine how well it performs. Just as in the medical setting, we are interested in the diagnostic's *False Discovery Rate (FDR)*, given by

$$FDR = A_{FP}/A_P = A_{FP}/(A_{FP} + A_{TP}), \tag{1.9}$$

and its *False Nondiscovery Rate (FNR)*, given by

$$FNR = A_{FN}/A_N = A_{FN}/(A_{TN} + A_{FN}). \tag{1.10}$$

Notice that the FNR and FDR are obtained from the first and second *rows*, respectively, of the 2×2 table given by Table 1.1.

In our evaluation of a diagnostic, we treat it as an algorithm that acts on the n spatial data and, for better or for worse, separates $Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n)$ into negatives and positives. A *summary* of this is captured by the counts, A_N and A_P (where recall $A_N + A_P = n$), but the full results of which datum is negative and which is positive are available and can be considered as part of the output of the algorithm. Hence, for a *given algorithm* (i.e., diagnostic), the *row totals* A_N and A_P of Table 1.1 *are given*. Consequently, our statistical evaluation is derived from the distribution of A_{FN} and A_{FP}, given A_N and A_P.

Several statistics are routinely used to assess the performance of medical diagnostics, and a similar approach can be used here for model diagnostics. The *Specificity*, or True negative rate, is

$$Sp \equiv A_{TN}/(A_{TN} + A_{FP}), \tag{1.11}$$

which is obtained from the first *column* of Table 1.1. The denominator of (1.11) is the number (out of n) that are in fact negative, as determined by the precise, follow-up reanalysis. In a hypothesis-testing setting, $1 - Sp$ is analogous to

$$\text{size} = \alpha \equiv \text{Type I error rate.}$$

The *Sensitivity*, or True positive rate, is

$$Se \equiv A_{TP}/(A_{FN} + A_{TP}), \tag{1.12}$$

which is obtained from the second *column* of Table 1.1. The denominator of (1.12) is the number (out of n) that are in fact positive, as determined by the precise, follow-up reanalysis. In a hypothesis-testing setting, Se is analogous to

$$\text{power} = 1 - \beta \equiv 1 - \text{Type II error rate.}$$

In Section 1.4, we suggest alternatives to Sp and Se for assessing the performance of model diagnostics. These are the False Discovery Rate (FDR) and the False Nondiscovery Rate (FNR) defined by (1.9) and (1.10), respectively.

Recall that we treat a model diagnostic as an algorithm that separates $Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n)$ into negatives and positives, and hence A_N and A_P in (1.8) are given. We propose that the precise, follow-up reanalysis of each spatial datum (to determine which of the negatives are True and which are False; and which of the positives are False and which are True) is obtained by *cross-validation* (e.g., Hastie et al. 2009, Section 7.10). The model diagnostic is based on a spatial model and the cross-validation is, of course, based on the *same* spatial model. It is worth noting that cross-validation is typically very slow to implement and, hence, we are only proposing to use it in evaluation. This is analogous to the way a cheap and easy medical diagnostic might be used in the general population, but its evaluation typically involves expensive but precise laboratory analysis.

For cross-validation in the spatial setting, a datum $Z(\boldsymbol{s}_i)$ is held out, and the spatial model is fitted to $\boldsymbol{Z}_{-i} \equiv (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_{i-1}), Z(\boldsymbol{s}_{i+1}), ..., Z(\boldsymbol{s}_n))^T$. That model is then used to predict $Z(\boldsymbol{s}_i)$ from data $\boldsymbol{Z}_{-i}$, resulting in a predictor of $Z(\boldsymbol{s}_i)$ that we denote $\hat{Z}_{-i}(\boldsymbol{s}_i)$. Then a negative at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{ll} \text{True if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq k_i \\ \text{False if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i; \end{array} \tag{1.13}$$

and a positive at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{False if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq k_i \\ \text{True if} & |\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i, \end{array} \tag{1.14}$$

where $\{k_i : i = 1, ..., n\}$ are thresholds determined by the variability in the cross-validation errors,

$$\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i); i = 1, ..., n. \tag{1.15}$$

Hence, given the negatives (whose number is A_N) and the positives (whose number is $A_P = n - A_N$), through (1.13) and (1.14) we can obtain all the numbers in Table 1.1. Consequently, we can compute the FDR given by (1.9), the FNR given by (1.10), the Sp given by (1.11) and the Se given by (1.12). We shall see in Section 1.4 how these quantities can be used to evaluate and compare spatial-model diagnostics. However, we first discuss the various entries in Table 1.1, for non-hierarchical models and then for hierarchical models.

1.3.1 Non-hierarchical spatial model

The concepts from which our diagnostic evaluation follows are clearest in the non-hierarchical case. Here, data $\boldsymbol{Z}$ are fitted directly to a spatial model without invoking a hidden model Y to deal with measurement error and "missingness." The original geostatistical paradigm (Matheron 1963) makes no distinction between $\boldsymbol{Z}$ and Y, and we start with this case. In a sense, this non-hierarchical spatial model is a special case of the hierarchical model in (1.1) and (1.2), where the data-model's error variance is zero (e.g., $\sigma_\delta^2 = 0$ for (1.5)). Then, at the location $\boldsymbol{s}_i$ where $Z(\boldsymbol{s}_i)$ is observed, the conditional distribution, $[Z(\boldsymbol{s}_i)|Y] = [Z(\boldsymbol{s}_i)|Z(\boldsymbol{s}_i)]$, is degenerate.

The missing data, which are at locations other than $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$, represent unknowns in the model. For example, if there is no observation at $\boldsymbol{s}_0$, then we wish to predict $Z(\boldsymbol{s}_0)$ given $\boldsymbol{Z}$. Kriging (e.g., Cressie 1993, Chapter 3) is based on this. Thus, in the non-hierarchical case, we wish to obtain $[Z(\boldsymbol{s}_0)|\boldsymbol{Z}]$, sometimes called the predictive distribution, to make inference on the missing datum $Z(\boldsymbol{s}_0)$. We shall see in Section 1.3.2 that this goal generalizes to wishing to obtain $[Y(\boldsymbol{s})|\boldsymbol{Z}]$, for all $\boldsymbol{s}$ in the spatial domain of interest.

Cross-validation means that $Z(\boldsymbol{s}_i)$ is predicted from $[Z(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}]$. That predictor was notated $\hat{Z}_{-i}(\boldsymbol{s}_i)$ above, and a common example is,

$$\hat{Z}_{-i}(\boldsymbol{s}_i) = E(Z(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}); \tag{1.16}$$

other predictors are possible. The cross-validation error (1.15) is substituted into (1.13) and (1.14) to determine which of the negatives and positives are True or False, and the counts are summarized in Table 1.1.

The SIDS example discussed in Section 1.2 involved two different diagnostics. The 2 × 2 table for each of them is given in Table 1.2 and Table 1.3. The threshold k_i used for location $\boldsymbol{s}_i$ is given by

$$k_i = k\sigma_\delta / N(\boldsymbol{s}_i)^{1/2}, \tag{1.17}$$

where k is chosen so that

$$\text{Pr}(|N(0,1)| \leq k) = \text{Pr}(|N(0,1)| \geq k) = 0.5,$$

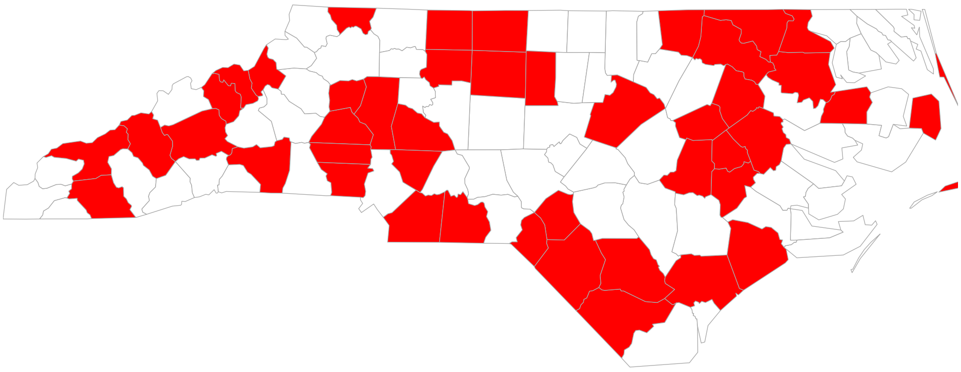

Figure 1.4 Cross-validation for the null model fitted to the transformed SIDS rates: Positive (i.e., unusually large) values are shaded.

Table 1.2 The 2×2 table given by Table 1.1, for the Local Moran I diagnostic applied to the transformed SIDS residuals after fitting the null model; cross-validation is abbreviated as CV.

	CV Negative	CV Positive	Total
Diagnostic Negative	54	28	82
Diagnostic Positive	2	16	18
Total	56	44	100

and N(0,1) is a standard normal random variable. This results in $k = 0.675$, which ensures that we give equal probability to being inside or outside the limit, assuming that the model fits. The map of positives given by cross-validation, namely the counties where $|\hat{Z}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > k_i$, for $i = 1, ..., n$, is shown in Figure 1.4.

Values of smaller k in (1.17) are of obvious interest because the precise, follow-up reanalysis is then very stringent; and values up to $k = 1.96$ satisfy $Pr(|N(0,1)| \leq k) \leq 0.95$. Hence we consider k to vary from small values near zero to values up to 2; in Section 1.4.1, it leads to a new type of curve that we call the *Discovery curve*.

Table 1.3 The 2×2 table given by Table 1.1, for the Local Getis-Ord G^* diagnostic applied to the transformed SIDS residuals after fitting the null model; cross-validation is abbreviated as CV.

	CV Negative	CV Positive	Total
Diagnostic Negative	53	35	88
Diagnostic Positive	3	9	12
Total	56	44	100

1.3.2 Hierarchical spatial model

From the hierarchical model (1.1) and (1.2), there is a hidden process $Y(\cdot)$ that is to be inferred. In this case, the cross-validation error is,

$$\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i); i = 1, ..., n, \tag{1.18}$$

where $\hat{Y}_{-i}(\boldsymbol{s}_i)$ is a predictor of $Y(\boldsymbol{s}_i)$ obtained from the predictive distribution, $[Y(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}]$. A common example is,

$$\hat{Y}_{-i}(\boldsymbol{s}_i) = E(Y(\boldsymbol{s}_i)|\boldsymbol{Z}_{-i}).$$

Ideally, we would like to base the criterion for True/False negatives/positives on the error, $\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i)$. However, $Y(\boldsymbol{s}_i)$ is unavailable.

In the hierarchical spatial model, (1.13) and (1.14) are modified, respectively, to: A negative at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{True if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq m_i \\ \text{False if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > m_i; \end{array} \tag{1.19}$$

and a positive at $\boldsymbol{s}_i$ is declared:

$$\begin{array}{rl} \text{False if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| \leq m_i \\ \text{True if} & |\hat{Y}_{-i}(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i)| > m_i. \end{array} \tag{1.20}$$

The threshold m_i used for location $\boldsymbol{s}_i$ is determined as follows: From (1.1) and (1.2), $Z(\boldsymbol{s}_i) = Y(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i)$, and hence the cross-validation error given by (1.18) is:

$$\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i) - \epsilon(\boldsymbol{s}_i),$$

where $\epsilon(\boldsymbol{s}_i)$ is independent of $Y(\boldsymbol{s}_i)$ and $\hat{Y}_{-i}(\boldsymbol{s}_i)$. Its variance is:

$$\text{var}(\hat{Y}_{-i}(\boldsymbol{s}_i) - Y(\boldsymbol{s}_i)) + \text{var}(\epsilon(\boldsymbol{s}_i)).$$

Thus, m_i is obtained in a similar manner to k_i with a modification to account for the measurement error, var$(\epsilon(\boldsymbol{s}_i)) \equiv \sigma_\epsilon(\boldsymbol{s}_i)^2$.

If a hierarchical model like that given by Cressie (1989) were fitted to the SIDS data in Section 1.2, we would have $\sigma_\epsilon(\boldsymbol{s}_i)^2 = \sigma_\epsilon^2/N(\boldsymbol{s}_i)$, and hence

$$\text{var}(Z(\boldsymbol{s}_i)) = (\sigma_\delta^2 + \sigma_\epsilon^2)/N(\boldsymbol{s}_i),$$

where we assume σ_ϵ^2 is known (e.g., from spatial-sampling considerations). Consequently, (1.17) is modified to give the following threshold in (1.19) and (1.20):

$$m_i = k(\sigma_\delta^2 + \sigma_\epsilon^2)^{1/2}/N(\boldsymbol{s}_i)^{1/2}, \tag{1.21}$$

where once again $k = 0.675$ gives equal probability to being inside or outside the limit, assuming that the model fits. By varying k from small values near 0 to values up to 2, a Discovery curve for the hierarchical spatial case is obtained; see Section 1.4.2.

1.4 Evaluation of Diagnostics

When evaluating medical diagnostics, biostatisticians often use the Receiver Operating Characteristic (ROC) curve (e.g., Metz 1978), which is a plot of Se (on the vertical axis) versus $1 - Sp$ (on the horizontal axis). It is well known from hypothesis testing that the Type I error rate (i.e., $1 - Sp$) and the Type II error rate (i.e., $1 - Se$) cannot both be kept small. Significance testing puts an upper bound on the Type I error rate (the level of significance) and uses tests whose $1-$Type II error rate is large (preferably maximized). To evaluate a medical diagnostic, it is recognized that Sp and Se will co-vary, which is captured by an (x, y) curve in $[0, 1] \times [0, 1]$, where

$$x = 1 - Sp \qquad \text{and} \qquad y = Se.$$

This defines an *ROC curve*, and ideally it is confined to a region of the domain that is close to $(x, y) = (0, 1)$, or at the very least it maintains a consistently high Se for most values of $1 - Sp$. Furthermore, two diagnostics can be compared using their respective ROC curves, by ascertaining which values of $1 - Sp$ lead to a uniformly dominant Se value for one diagnostic over the other. A definitive ordering of several medical diagnostics can be obtained through the *areas* under their respective ROC curves (e.g., Fawcett 2006). In Table 1.1, the ROC curve computes rates with respect to each *column* and plots them. Craiu and Sun (2008) propose another type of curve with $x = FDR$ and $y = 1 - Se$, which involves error rates from both a row and a column of Table 1.1.

When a medical diagnostic is applied many times over, error rates computed with respect to the two *rows* of the 2×2 table are more relevant. The analogy to spatial-model diagnostics is immediate, where each datum $Z(\boldsymbol{s}_i)$ at spatial location $\boldsymbol{s}_i$, for $i = 1, ..., n$, is potentially a positive or a negative. Thus, we propose to replace the ROC curve with something we call a *Discovery (DSC) curve*; it is an (x, y) curve in $[0, 1] \times [0, 1]$, where

$$x = FDR \qquad \text{and} \qquad y = 1 - FNR,$$

for FDR and FNR given by (1.9) and (1.10), respectively.

The DSC curve captures the rate of False positives among all positives (plotted on the x-axis) and the rate of True negatives among all negatives (plotted on the y-axis). Ideally, the curve is confined to a region of the domain that is close to $(x, y) = (0, 1)$, or at the very least it maintains a consistently high $1 - FNR$ for most values of FDR. Hence, two diagnostics for a spatial model can be compared using their respective DSC curves, and a definitive ordering can be obtained through the areas under their respective curves.

In the next two subsections, we pursue the DSC-curve approach to evaluating diagnostics, first for non-hierarchical spatial models and then for hierarchical spatial models.

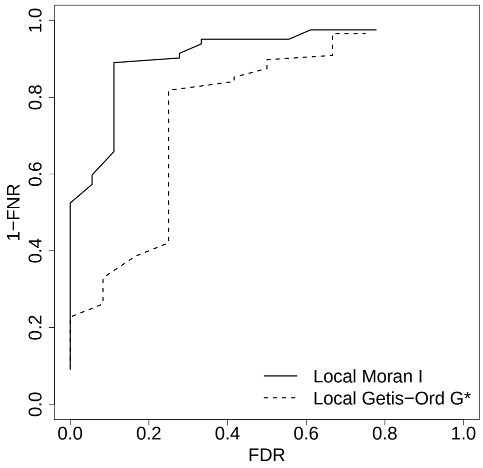

Figure 1.5 DSC curves for the SIDS data, for $0 < k < 2$ in (1.17).

1.4.1 DSC curves for non-hierarchical spatial models

Table 1.1 is obtained from (1.13) and (1.14). If each entry in the table is seen as a function of $\boldsymbol{k} = (k_1, ..., k_n)^T$, then by varying $\boldsymbol{k}$ a DSC curve can be obtained. The SIDS example discussed in Section 1.2 and above in this section, has a 2×2 table that is determined by a single, normalized threshold k; see (1.17). By varying k from near 0 up to 2, we obtain a DSC curve for each of the two diagnostics. These are shown in Figure 1.5.

Recall the interpretation of these DSC curves; Figure 1.5 shows uniformly superior behaviour of the local Moran I diagnostic compared with the local Getis-Ord G^* diagnostic.

1.4.2 DSC curves for hierarchical spatial models

Because a DSC curve depends on Table 1.1, if we can find such a 2×2 table for a hierarchical spatial model, then everything proceeds as in Section 1.4.1. From Section 1.3.2, we see that each entry in the 2×2 table can be seen as a function of the thresholds $\boldsymbol{m} = (m_1, ..., m_n)^T$. Then by varying $\boldsymbol{m}$ a DSC curve can be obtained.

If a hierarchical model like that given by Cressie (1989) were fitted to the SIDS data in Section 1.2, we have seen in Section 1.3.2 that $\boldsymbol{m}$ would depend only on a single k (equation (1.21)) that could be varied from small values near 0 to values up to 2. This would result in a DSC curve for the hierarchical spatial model fitted to the SIDS data, representing the next step in this line of research.

1.5 Discussion and Conclusions

This chapter explores the strong analogy between medical diagnostics and spatial-hierarchical-model diagnostics. A spatial datum is analogous to an individual whose health is being diagnosed. Medical diagnostics can be evaluated with ROC curves, and in some applications they are investigated using the concept of False Discovery Rate. We have made the observation that a different curve, which we have called the Discovery (DSC) curve, gives another way to evaluate a diagnostic. For a spatial model, the True negatives and False positives are defined in our proposed evaluation procedure through cross-validation.

By its very nature, a spatial model describes statistical dependence between the data $\boldsymbol{Z}$. Hence, the cross-validation errors given by (1.15) or (1.18) are themselves spatially dependent. In future research, we wish to go beyond our descriptive, visual evaluation of a spatial-model diagnostic and address questions like, “What is the confidence region for a given $(E(FDR), E(1 - FNR))$ pair?” and “Are two DSC curves significantly different?”

Cross-validation is almost always computationally expensive, which is why other diagnostics are preferred when datasets are massive. In this work on evaluation of a model diagnostic, we are willing to spend the computing resources to gauge a diagnostic's “goodness” on benchmark datasets.

Cross-validation is just one way to define a precise, follow-up reanalysis that is used to determine the counts in Table 1.1. Another way would be to base this reanalysis on “testing datasets” proposed by Efron (1983, 1986), which adapt well to the hierarchical-model setting.

Instead of Table 1.1 for non-hierarchical models, this chapter is really about a $2 \times 2 \times 2$ table for hierarchical models where the extra dimension captures a 2×2 table for the Z-process on top of a 2×2 table for the Y-process. The bottom table is hidden since Y is hidden, but it could be thought of as representing an “oracle” table. In this chapter, we have given ways to construct an appropriate 2×2 table and hence an appropriate DSC curve that recognizes the hierarchical nature (i.e., presence of a hidden process Y) of the spatial model, without appealing to the oracle table.

Acknowledgements

This research was partially supported by the U.S. National Science Foundation (NSF) and the U.S. Census Bureau under NSF grant SES-1132031, funded through the NSF-Census Research Network (NCRN) program. The authors would like to express their appreciation to the editors for their helpful comments.

References

Akobeng AK 2007 Understanding diagnostic tests 1: sensitivity, specificity and predictive values. *Acta paediatrica* **96**, 338–341.

Anselin L 1995 Local indicators of spatial association—LISA. *Geographical Analysis* **27**, 93–115.

Anselin L and Rey SJ 2010 *Perspectives on Spatial Data Analysis*. Springer, Heidelberg and New York, NY.

Baddeley A, Turner R, Møller J and Hazelton M 2005 Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society, Series B* **67**, 617–666.

Banerjee S, Carlin BP and Gelfand AE 2004 *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL.

Bayarri MJ and Berger JO 2000 P-values for composite null models. *Journal of the American Statistical Association* **95**, 1127–1142.

Bayarri MJ and Castellanos ME 2007 Bayesian checking of the second levels of hierarchical models. *Statistical Science* **22**, 322–343.

Belsley DA, Kuh E and Welsch RE 1980 *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, NY.

Benjamini Y and Heller R 2007 False discovery rates for spatial signals. *Journal of the American Statistical Association* **102**, 1272–1281.

Benjamini Y and Hochberg Y 1995 Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini Y and Hochberg Y 1997 Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

Benjamini Y and Yekutieli D 2001 The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Bivand R 2014 *spdep: Spatial dependence: weighting schemes, statistics and models.* R package version 0.5-74.

Boots B 2003 Developing local measures of spatial association for categorical data. *Journal of Geographical Systems* **5**, 139–160.

Bousquet N 2008 Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics* **35**, 1011–1029.

Box GEP 1980 Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A* **143**, 383–430.

Bradley R and Haslett J 1992 High-interaction diagnostics for geostatistical models of spatially referenced data. *The Statistician* **41**, 371–380.

Carlin BP and Louis TA 2009 *Bayesian Methods for Data Analysis* 3rd edn. Chapman and Hall/CRC, Boca Raton, FL.

Christensen R, Johnson W and Pearson LM 1992 Prediction diagnostics for spatial linear models. *Biometrika* **79**, 583–591.

Cook RD and Weisberg S 1982 *Residuals and Influence in Regression*. Chapman and Hall, New York, NY.

Cox DR and Snell EJ 1968 A general definition of residuals. *Journal of the Royal Statistical Society, Series B* **30**, 248–275.

Craiu RV and Sun L 2008 Choosing the lesser evil: trade-off between false discovery rate and non-discovery rate. *Statistica Sinica* **18**, 861 – 879.

Crespi CM and Boscardin WJ 2009 Bayesian model checking for multivariate outcome data. *Computational Statistics and Data Analysis* **53**, 3765–3772.

Cressie N 1989 Empirical Bayes estimation of undercount in the Decennial Census. *Journal of the American Statistical Association* **84**, 1033–1044.

Cressie N 1993 *Statistics for Spatial Data* rev. edn. John Wiley and Sons, New York, NY.

Cressie N and Chan NH 1989 Spatial modeling of regional variables. *Journal of the American Statistical Association* **84**, 393–401.

Cressie N and Read TRC 1985 Do sudden infant deaths come in clusters?. *Statistics and Decisions* **Supplement Issue 2**, 333–349.

Cressie N and Wikle C 2011 *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ.

Dahl FA 2006 On the conservativeness of posterior predictive p-values. *Statistics and Probability Letters* **76**, 1170–1174.

Dey D, Gelfand A, Swartz T and Vlachos P 1998 A simulation-intensive approach for checking hierarchical models. *Test* **7**, 325–346.

Dormann CF, McPherson JM, Araújo MB, Bivand R, Bolliger J, Carl G, Davies RG, Hirzel A, Jetz W, Kissling WD, Kühn I, Ohlemüller R, Peres-Neto PR, Reineking B, Schröder B, Schurr FM and Wilson R 2007 Methods to account for spatial autocorrelation in the analysis of species distributional data: A review. *Ecography* **30**, 609–628.

Efron B 1983 Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron B 1986 How biased is the apparent error rate of a prediction rule?. *Journal of the American Statistical Association* **81**, 461–470.

Efron B 2004 The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* **99**, 619–642.

Enøe C, Georgiadis MP and Johnson WO 2000 Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. *Preventive Veterinary Medicine* **45**, 61–81.

Evans M and Moshonov H 2006 Checking for prior-data conflict. *Bayesian Analysis* **1**, 893–914.

Fawcett T 2006 An introduction to ROC analysis. *Pattern Recognition Letters* **27**, 861–874.

Finner H, Dickhaus T and Roters M 2007 Dependency and false discovery rate: Asymptotics. *Annals of Statistics* **35**, 1432–1455.

Fotheringham AS 2009 The problem of spatial autocorrelation and local spatial statistics. *Geographical Analysis* **41**, 398–403.

Fotheringham AS and Brunsdon C 1999 Local forms of spatial analysis. *Geographical Analysis* **31**, 340–358.

Fox J 1991 *Regression Diagnostics*. Sage Publications, Newbury Park, CA.

Gelfand AE 1996 Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (ed. Gilks WR, Richardson S and Spiegelhalter DJ) Chapman & Hall, London, UK pp. 145–161.

Gelfand AE, Dey DK and Chang H 1992 Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4* (ed. Bernardo JM, Berger JO, Dawid AP and Smith A) Oxford University Press, Oxford, UK pp. 147–167.

Gelfand AE, Diggle PJ, Fuentes M and Guttorp, P. (ed.) 2010 *Handbook of Spatial Statistics*. Chapman and Hall/CRC, Boca Raton, FL.

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A and Rubin DB 2013 *Bayesian Data Analysis* 3rd edn. Chapman and Hall/CRC, Boca Raton, FL.

Gelman A, Meng XL and Stern HS 1996 Posterior predictive asssessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Genovese C and Wasserman L 2002 Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B* **64**, 499–517.

Getis A and Ord JK 1992 The analysis of spatial association by use of distance statistics. *Geographical Analysis* **24**, 189–206.

Glatzer E and Müller WG 2004 Residual diagnostics for variogram fitting. *Computers and Geosciences* **30**, 859–866.

Gneiting T 2011 Making and evaluating point forecasts. *Journal of the American Statistical Association* **106**, 746–762.

Goel PK and De Groot MH 1981 Information about hyperparameters in hierarchical models. *Journal of the American Statistical Association* **76**, 140–147.

Hastie T, Tibshirani R and Friedman JH 2009 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd edn. Springer, New York, NY.

He D, Xu X and Liu X 2013 The use of posterior predictive p-values in testing goodness-of-fit. *Communications in Statistics - Theory and Methods* **42**, 4287–4297.

Hering AS and Genton MG 2011 Comparing spatial predictions. *Technometrics* **53**, 414–425.

Hill SD and Spall JC 1994 Sensitivity of a Bayesian analysis to the prior distribution. *IEEE Transactions on Systems, Man and Cybernetics* **24**, 216–221.

Hjort NL, Dahl FA and Steinbakk GH 2006 Post-processing posterior predictive p-values. *Journal of the American Statistical Association* **101**, 1157–1174.

Hu JX, Zhao H and Zhou HH 2010 False discovery rate control with groups. *Journal of the American Statistical Association* **105**, 1215–1227.

Huber-Carol C, Balakrishnan N, Nikulin MS and Mesbah M 2002 *Goodness-of-Fit Tests and Model Validity*. Birkhäuser, Boston, MA.

Hui SL and Zhou XH 1998 Evaluation of diagnostic tests without gold standards. *Statistical Methods in Medical Research* **7**, 354–370.

Kaiser MS, Lahiri SN and Nordman DJ 2012 Goodness-of-fit tests for a class of Markov random field models. *Annals of Statistics* **40**, 104–130.

Karlström A and Ceccato V 2002 A new information theoretical measure of global and local spatial association. *Jahrbuch für Regionalwissenschaft* **22**, 13–40.

Kulldorff M, Huang L, Pickle L and Duczmal L 2006 An elliptic spatial scan statistic. *Statistics in Medicine* **25**, 3929–3943.

Le Rest K, Pinaud D, Monestiez P, Chadoeuf J and Bretagnolle V 2014 Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography* **23**, 811–820.

Lee H and Ghosh SK 2009 Performance of information criteria for spatial models. *Journal of Statistical Computation and Simulation* **79**, 93–106.

Loy A and Hofmann H 2013 Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics* **5**, 48–61.

Marshall EC and Spiegelhalter DJ 2003 Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine* **22**, 1649–1660.

Marshall EC and Spiegelhalter DJ 2007 Identifying outliers in Bayesian hierarchical models: A simulation-based approach. *Bayesian Analysis* **2**, 409–444.

Massmann C, Wagener T and Holzmann H 2014 A new approach to visualizing time-varying sensitivity indices for environmental model diagnostics across evaluation time-scales. *Environmental Modelling and Software* **51**, 190–194.

Matheron G 1963 Principles of geostatistics. *Economic Geology* **58**, 1246–1266.

Meng XL 1994 Posterior predictive p-values. *Annals of Statistics* **22**, 1142–1160.

Metz CE 1978 Basic principles of ROC analysis. *Seminars in Nuclear Medicine* **8**, 283–298.

Moraga P and Montes F 2011 Detection of spatial disease clusters with LISA functions. *Statistics in Medicine* **30**, 1057–1071.

Murray K, Heritier S and Müller S 2013 Graphical tools for model selection in generalised linear models. *Statistics in Medicine* **32**, 4438–4451.

O'Hagan A 2003 HSSS model criticism In *Highly Structured Stochastic Systems* (ed. Green PJ, Hjort NL and Richardson S) Oxford University Press, Oxford, UK pp. 423–443.

Ord JK and Getis A 1995 Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* **27**, 286–306.

Ord JK and Getis A 2012 Local spatial heteroscedasticity (LOSH). *Annals of Regional Science* **48**, 529–539.

Pepe MS and Thompson ML 2000 Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

Presanis AM, Ohlssen D, Spiegelhalter DJ and de Angelis D 2013 Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statistical Science* **28**, 376–397.

R Core Team 2014 *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.

Read S, Bath PA, Willett P and Maheswaran R 2013 New developments in the spatial scan statistic. *Journal of Information Science* **39**, 36–47.

Robertson C, Long JA, Nathoo FS, Nelson TA and Plouffe CCF 2014 Assessing quality of spatial models using the structural similarity index and posterior predictive checks. *Geographical Analysis* **46**, 53–74.

Robins JM, Van der Vaart A and Ventura V 2000 Asymptotic distribution of p-vlues in composite null models. *Journal of the American Statistical Association* **95**, 1143–1156.

Sackett DL and Haynes RB 2002 The architecture of diagnostic research. *British Medical Journal* **324**, 539–541.

Schabenberger O and Gotway CA 2005 *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

Scheel ID, Green PJ and Rougier JC 2011 A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models. *Scandinavian Journal of Statistics* **38**, 529–550.

Sengupta A and Cressie N 2013 Empirical hierarchical modelling for count data using the spatial random effects model. *Spatial Economic Analysis* **8**, 389–418.

Shen X, Huang HC and Cressie N 2002 Nonparametric hypothesis testing for a spatial signal. *Journal of the American Statistical Association* **97**, 1122–1140.

Steinbakk GH and Storvik GO 2009 Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics* **36**, 320–336.

Stern HS and Cressie N 2000 Posterior predictive model checks for disease mapping models. *Statistics in Medicine* **19**, 2377–2397.

Stone M 1974 Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B* **36**, 111–147.

Storey JD 2003 The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* **31**, 2013–2035.

Storey JD and Tibshirani R 2003 Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440–9445.

Symons MJ, Grimson RC and Yuan YC 1983 Clustering of rare events. *Biometrics* **39**, 193.

van Smeden M, Naaktgeboren CA, Reitsma JB, Moons KGM and de Groot JAH 2014 Latent class models in diagnostic studies when there is no reference standard - a systematic review. *American Journal of Epidemiology* **179**, 423–431.

Wang Z, Bovik AC, Sheikh HR and Simoncelli EP 2004 Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* **13**, 600–612.

Xu M, Mei CL and Yan N 2014 A note on the null distribution of the local spatial heteroscedasticity (LOSH) statistic. *Annals of Regional Science* **52**, 697–710.

Yan G and Sedransk J 2007 Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Analysis* **2**, 735–760.

Yuan Y and Johnson VE 2012 Goodness-of-fit diagnostics for Bayesian hierarchical models. *Biometrics* **68**, 156–164.

Zhang H and Wang Y 2010 Kriging and cross-validation for massive spatial data. *Environmetrics* **21**, 290–304.