4-2007

# Exploring Dependence with Data on Spatial Lattices

Mark S. Kaiser
*Iowa State University*, mskaiser@iastate.edu

Petruta C. Caragea
*Iowa State University*, pcaragea@iastate.edu

# Exploring Dependence with Data on Spatial Lattices

**Abstract**

The application of Markov random field models to problems involving spatial data on lattice systems requires decisions regarding a number of important aspects of model structure. Existing exploratory techniques appropriate for spatial data do not provide direct guidance to an investigator about these decisions. We introduce an exploratory quantity that is directly tied to the structure of Markov random field models based on one parameter exponential family conditional distributions. This exploratory diagnostic is shown to be a meaningful statistic that can inform decisions involved in modeling spatial structure with statistical dependence terms. In this article, we develop the diagnostic, show that it has stable statistical behavior, illustrate its use in guiding modeling decisions with simulated examples, and demonstrate that these properties have use in applications.

**Keywords**

Auto-models, conditionally specified models, spatial structure, statistical dependence, exploratory analysis

**Disciplines**

Statistics and Probability

**Comments**

This preprint was published as Mark Kaiser and Petruta Caragea, "Exploring Dependence with Data on Spatial Lattices" *Biometrics* (2009): 857-865, doi: 10.1111/j.1541-0420.2008.01118.x.

# Exploring Dependence with Data on Spatial Lattices

Mark S. Kaiser and Petruţa C. Caragea

Department of Statistics

Iowa State University

April 2007

# Summary

The application of Markov random field models to problems involving spatial data on lattice systems requires decisions regarding a number of important aspects of model structure. Existing exploratory techniques appropriate for spatial data do not provide direct guidance to an investigator about these decisions. We introduce an exploratory quantity that is directly tied to the structure of Markov random field models based on one parameter exponential family conditional distributions. This exploratory diagnostic is shown to be a meaningful statistic that can inform decisions involved in modeling spatial structure with statistical dependence terms. In this article, we develop the diagnostic, show that it has stable statistical behavior, illustrate its use in guiding modeling decisions with simulated examples, and demonstrate that these properties have use in applications.

KEYWORDS: Auto-models, conditionally specified models, spatial structure, statistical dependence, exploratory analysis.

# 1 Introduction

The application of Markov random field models to spatial problems in the environmental and ecological sciences, agriculture, and other areas of biology is growing. Along with models based on observable random variables having conditional normal distributions (see, e.g., Haining, 1990; Griffith and Layne, 1999; Rue and Held, 2005), binary Markov random field models have been used to model, among other problems, spatial patterns of plant disease (Gumpertz, Graham and Ristaino, 1997), spatial distributions of plant species (Wu and Huffer, 1997), and outbreaks of southern pine beetle (Zhu, Huang and Wu, 2005). Models with Winsorized Poisson conditionals or truncated Poisson conditionals have been applied to modeling counts of agricultural and environmental variables (e.g., Augustin, McNicol and Marriot, 2006; Kaiser, 2001). Markov random field models are also now commonly used in hierarchical models, and present one option for incorporating both time and space effects (e.g., Wikle, Milliff, Nychka and Berliner, 2001; Rue and Held, 2005; Kaiser, Daniels, Furukawa and Dixon, 2002). While Markov random field models are applicable to a wider variety of settings (e.g., Kaiser, 2001), spatial data on a regular lattice provide a natural context within which to consider questions involved in model formulation and analysis. We will restrict ourselves to this situation in what is to follow.

The problems addressed by various authors in the preceding references illustrate that the formulation of Markov random field models requires, *inter alia*, specification of a neighborhood for each location, choice of parameterizations to represent unidirectional, directional, or possibly even time-varying dependencies, and the decision to include or not include covariates. Typically, such modeling choices depend on fitting models of differing structures and examining model performance based on criteria such as penalized likelihood values or predictive ability (e.g., Gumpertz *et al.*, 1997; Hoeting, Leecaster and Bowden, 2000; Kaiser *et al.*, 2002). It would be beneficial in practice to have exploratory tools to help guide some of the modeling choices that must be made in the analysis of a given problem.

Any number of basic summary statistics are available to detect spatial structure, such as join statistics for binary data, Moran's I statistic for more general variables, and various forms of sample autocorrelation functions (e.g., Schabenberger and Gotway, 2005). These statistics can provide indications of spatial structure in sets of data, but they do not provide an indication of how one might go about modeling that structure. The classical variogram of geostatistical analysis comes closer to providing useful modeling information through the structures of nugget, range and sill (e.g., Cressie, 1993). Variograms are also commonly used to examine data for indications that modeling directional structures might be profitable. But it is not clear how the components of typical Markov random field models are related to these characteristics of variograms (although see Rue and Held, 2005, Chapter 5 for discussion in terms of covariance functions with Gaussian fields). This motivates a desire to develop an exploratory tool that is tied more specifically to the parametric structure of particular models. The primary objectives of this article are to propose one such exploratory quantity, demonstrate that it has reasonable statistical behavior, and illustrate several potential uses with simulated data and previously published analyses.

Throughout this article we will use the terms *statistical dependence* and *spatial structure*, rather than the more generic term spatial dependence. Although easily misunderstood, it is known (Cressie, 1993, p. 114) that what is often thought of as spatial dependence may sometimes be represented by spatial trend in a model with independent random variables, may sometimes be represented by a model with no spatial trend but random variables having dependencies that are functions of spatial location, and may sometimes be represented by a model having both trend and dependence components. It is a modeling choice as to what representation is best suited for a particular problem. Our concern is in modeling statistically dependent random variables and we will refer to spatial structure as inclusive of patterns produced from either trend or dependence components, reserving the term statistical dependence for model components relating to non-independent random variables.

The remainder of the article is organized as follows. In Section 2 we develop an exploratory diagnostic we call the S-value. Simulations are presented in Section 3 to examine the behavior of the S-value for several simple models. Section 4 extends the basic S-value formulation to situations involving directional dependence parameters and spatial trend (including covariates). In Section 5 we present several simulated examples that demonstrate the usefulness of the S-value in detecting strength of dependence, directional dependence, and spatial trend. A re-examination of several published applications of Markov random field models is contained in Section 6, and concluding remarks are offered in Section 7.

## 2 A Model Based Exploratory Diagnostic

While not essential for our development, we will restrict attention to two-dimensional real space and assume there are available a set of spatial locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ on a regular square lattice, where $\boldsymbol{s}_i \equiv (u_i, v_i)$ denotes a location at horizontal coordinate $u_i$ and vertical coordinate $v_i$. We also assume that each location has a designated neighborhood $N_i \equiv \{\boldsymbol{s}_j : \boldsymbol{s}_j \text{ is a neighbor of } \boldsymbol{s}_i\}$. The simple configurations corresponding to four-nearest and eight-nearest neighbors are well suited for use with regular lattices and we will use those neighborhood structures repeatedly. We will also assume that each location has the same number of neighbors $m$, so that the situations being considered are either the theoretical lattice on a torus or, more relevant to practical applications, on a lattice with a border.

## 2.1 Exponential Family Markov Random Field Models

Given locations and neighborhoods, formulation of a Markov random field model involves specifying, for $i = 1, \ldots, n$, a full conditional probability mass or density function, assumed to depend functionally only on values at neighboring locations, $\boldsymbol{y}(N_i) \equiv \{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\}$. The models we will consider can be written in terms of one-parameter exponential families of the form,

$$f\{y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i)\} = \exp\left[A_i(\boldsymbol{y}(N_i))\, y(\boldsymbol{s}_i) - B_i(\boldsymbol{y}(N_i)) + C(y(\boldsymbol{s}_i))\right],$$

with, for equal valued dependence of a location with all of its neighbors,

$$A_i(\boldsymbol{y}(N_i)) = \tau^{-1}(\kappa_i) + \gamma \frac{1}{m} \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \kappa_j\}. \tag{1}$$

Specifically, we will consider models that have Gaussian conditionals and constant conditional variance $\sigma^2$ for which,

$$A_i(\boldsymbol{y}(N_i)) = \kappa_i/\sigma^2 + \gamma \frac{1}{m} \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \kappa_j\}, \tag{2}$$

binary conditionals, for which,

$$A_i(\boldsymbol{y}(N_i)) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \gamma \frac{1}{m} \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \kappa_j\}, \tag{3}$$

and Winsorized Poisson conditionals, for which,

$$A_i(\boldsymbol{y}(N_i)) = \log(\kappa_i) + \gamma \frac{1}{m} \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \kappa_j\}. \tag{4}$$

The parameterizations in expressions (1), (2), (3) and (4) are important, and represent centered versions of these models. For the Gaussian model of expression (2), the constants $\kappa_i$; $i = 1, \ldots, n$ are exactly the marginal expectations $E\{Y(\boldsymbol{s}_i)\}$ (e.g., Besag, 1974; Cressie, 1993). Furukawa (2004) demonstrates through Monte Carlo simulation that the centered models for binary and Winsorized Poisson conditionals have marginal expectations that are nearly equal to the parameters $\kappa_i$ in (3) and (4) for certain ranges of $\gamma$. Caragea and Kaiser (2007) discuss the effects of centering on interpretation and stability of large-scale and small-scale model parameters for binary conditionals or auto-logistic models, and Kaiser (2007) demonstrates the same phenomenon as Furukawa (2004) with exact computations for a binary conditionals model on a spatial transect. We will revisit the issue of how large $\gamma$ can be while still having the $\kappa_i$ in (2), (3), and (4) represent nearly marginal means in Section 2.4, as this is important for our interpretation of the exploratory quantity developed in the next subsection.

The Gaussian conditionals model in (2) is typically written in terms of conditional expectations, but we have given it here in terms of the natural parameter function $A_i$ for consistency with the other two models considered. Also, we follow the convention of Kaiser (2007) in writing the dependence parameter $\gamma$ in each model as a factor of the average discrepancy of neighboring values from their independence model expectations $\kappa_j$, rather than the sum. This facilitates interpretation of values of $\gamma$ across models with different neighborhood structures.

## 2.2 Development of the Proposed Measure

Exploratory measures for spatial structure are generally based on moment-like estimates of some quantification of statistical dependence. Sample variograms use moment-based estimates of the variance of differences. Moran's I statistic uses moment estimates of pairwise covariance among neighbors, as do autocorrelation measures. The concept of dependence embodied in exponential family Markov random field models is the discrepancy between expectations conditional on neighboring values, and expectations in the absence of spatial dependence. More precisely, the models "capture" spatial dependence as quantified by,

$$E\{Y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i)\} - E\{Y(\boldsymbol{s}_i)|\emptyset\} = \tau\{A_i(\boldsymbol{y}(N_i))\} - \kappa_i = \gamma\frac{1}{m}\sum_{\boldsymbol{s}_j \in N_i}\{y(\boldsymbol{s}_j) - \kappa_j\}\,, \quad (5)$$

where $E\{Y(\boldsymbol{s}_i)|\emptyset\}$ denotes expectation under a model with no statistical dependence (i.e., $\gamma = 0$). See Kaiser (2007) for extensive discussion of this concept of dependence in Markov random field models. A direct moment estimator of (5) is not readily available. But we can construct a quantity that reflects this difference through moment estimates of marginal and conditional means, which are available.

Consider a simple version of model (1), in which $\kappa_i = \kappa$ for all $i = 1, \ldots, n$. Then the parameter $\kappa$ is the common expectation of all $Y(\boldsymbol{s}_i)$ if these random variables are modeled as independent, and is also (nearly) the common marginal expectation under the dependence model, if $|\gamma|$ is less than its "standard bound", a concept that will be explained further in Section 2.4. The sample mean $\tilde{\kappa} = (1/n)\sum_i Y(\boldsymbol{s}_i)$ is then a moment estimator of $\kappa$. Recall that we are assuming that all locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ have the same number of neighbors $m$ so that there are no edge effects to account for. In any of the models considered here the neighboring values $\boldsymbol{y}(N_i)$ influence the conditional distribution only through the average $w(\boldsymbol{s}_i) = (1/m)\sum_{\boldsymbol{s}_j \in N_i} y(\boldsymbol{s}_j)$. Suppose that $w(\boldsymbol{s}_i);\ \ i = 1, \ldots, n$ can assume only values in a discrete, finite set as $w(\boldsymbol{s}_i) \in \{h_1, \ldots, h_q\}$ for $i = 1, \ldots, n$. This will be true for conditionals with discrete, finite support such as binary, binomial, and Winsorized Poisson cases, and can be produced in any situation by binning the sums of neighboring values in a particular data set, which we discuss in Section 2.3. Under these conditions, the natural parameter function $A_i$ of expression (1) becomes, for $\ell = 1, \ldots, q$,

$$A_i(\boldsymbol{y}(N_i)|w(\boldsymbol{s}_i) = h_\ell) = A_i(h_\ell) = \tau^{-1}(\kappa) + \gamma\left\{h_\ell - \kappa\right\},$$

or,

$$\frac{1}{\gamma}\left\{A_i(h_\ell) - \tau^{-1}(\kappa)\right\} = h_\ell - \kappa, \tag{6}$$

and moment estimates of these quantities are,

$$D(h_\ell, \tilde{\kappa}) = h_\ell - \tilde{\kappa}; \quad \ell = 1, \ldots, q. \tag{7}$$

Now, the conditional expectation of $Y(\boldsymbol{s}_i)$ given its neighboring values $\boldsymbol{y}(N_i)$ is $E\{Y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i)\} = \tau\left(A_i(\boldsymbol{y}(N_i))\right)$ which is again only influenced by the neighboring values through the average $w(\boldsymbol{s}_i)$. Let $H_\ell \equiv \{\boldsymbol{s}_i : w(\boldsymbol{s}_i) = h_\ell\}$; $\ell = 1, \ldots, q$. Then a moment estimator of $E\{Y(\boldsymbol{s}_i)|w(\boldsymbol{s}_i) = h_\ell\}$ is, for $\ell = 1, \ldots, q$,

$$C_\ell = \frac{1}{|H_\ell\,|} \sum_{\boldsymbol{s}_i \in H_\ell} Y(\boldsymbol{s}_i), \tag{8}$$

where $|H_\ell\,|$ denotes the number of locations in the set $H_\ell$. Then one would expect that $\{A_i(h_\ell) - \tau^{-1}(\kappa)\}$ in expression (6) should behave in a manner similar to the quantities,

$$r(C_\ell,\, \tilde{\kappa}) = \tau^{-1}(C_\ell) - \tau^{-1}(\tilde{\kappa}), \tag{9}$$

although these are not moment estimators because of the nonlinear transformations involved.

Finally, if $D(h_\ell, \tilde{\kappa})$ estimates the scaled difference in natural parameters between dependence and independence models, namely $(1/\gamma)\,\{A_i(h_\ell) - \tau^{-1}(\kappa)\}$, and $r(C_\ell,\, \tilde{\kappa})$ "estimates" the unscaled difference as immediately above, then we should have, for $\ell = 1, \ldots, q$, that $r(C_\ell,\, \tilde{\kappa}) \approx \gamma D(h_\ell, \tilde{\kappa})$. The proposed measure, which we call the S-value, is then given as the ordinary least squares estimate of slope for a regression through the origin of the $r(\cdot)$ on the $D(\cdot)$, namely,

$$S = \frac{\displaystyle\sum_{\ell=1}^{q} r(C_\ell,\, \tilde{\kappa}) D(h_\ell, \tilde{\kappa})}{\displaystyle\sum_{\ell=1}^{q} \{D(h_\ell, \tilde{\kappa})\}^2}. \tag{10}$$

In many cases there is either not a finite set of possible values for neighboring sums, such as for Gaussian conditionals, or the set is larger than can be expected to produce sufficient replicates for each value, such as for Winsorized Poisson conditionals. In these cases we propose the use of a data-driven binning procedure to construct the set of values $\{h_\ell : \ell = 1, \ldots q\}$ for use in calculating the S-value. Given observations $\{y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ and neighborhoods $\{N_i : i = 1, \ldots, n\}$, the values of $h_\ell$; $\ell = 1, \ldots, q$ can be computed by dividing the empirical distribution of the observed neighborhood averages $w(\boldsymbol{s}_i) = (1/m)\sum_{\boldsymbol{s}_j \in N_i} y(\boldsymbol{s}_j)$; $i = 1, \ldots, n$ into $q$ bins based on $q + 1$ quantiles (including the 0 and 1 quantiles). The values of $h_\ell$; $\ell = 1, \ldots, q$ are then set to the midpoints of these bins. See Web Appendix A for notational details defining the binning procedure in a formal manner. Note that binning implies that several possible sets of the $h_\ell$ and hence also several possible sets of the estimated conditional means $C_\ell$ are available for the same set of data. As long as there are a sufficient number of observations contributing to each bin, this should not have a major effect on the resultant value of $S$ in (10).

## 2.3 Standard Bounds

As mentioned previously, and will be illustrated in the sections to come, interpretation of S-values often hinges on their magnitudes relative to what Kaiser (2007) has proposed as "standard bounds" for values of the dependence parameter $\gamma$. In essence, standard bounds are limits on the allowable size of $|\gamma| < \gamma_{sb}$ in order that the leading constant terms in (1) can be interpreted as (nearly) the marginal means of $Y(\boldsymbol{s}_i)$; $i = 1, \ldots, n$. Continue, at this point, to consider the case $\kappa_i = \kappa$; $i = 1, \ldots, n$. For a model with Gaussian conditionals, the standard bound is $\gamma_{sb} = 1/\sigma^2$, independent of $\kappa$. For models with Winsorized Poisson conditionals and Winsorization value $R$, the standard bound is $\gamma_{sb} = \{\log(R) - \log(\kappa)\}/(R - \kappa)$, depending on both $R$ and $\kappa$. For models with binary conditionals, $\gamma_{sb}$ must be determined numerically, and depends on $\kappa$. As demonstrated by Kaiser (2007), one can also form what might be called "uniform standard bounds" that apply to any possible value of $\kappa$. For Gaussian models this uniform bound is the same as the standard bound, $1/\sigma^2$. For Winsorized Poisson models the uniform bound is $\gamma_u = 1/R$, and for binary models the value is a constant $\gamma_u = 4.0$. There is not much difference between standard bounds and the uniform standard bound for binary models unless $\kappa$ becomes extreme (such as $\kappa < 0.1$ or $\kappa > 0.9$) but for Winsorized Poisson models the difference between standard and uniform standard bounds is important in allowing for strong levels of statistical dependence.

Standard bounds are not "sharp" in the sense of being required for a valid model, that is, a model for which a joint distribution exists, although this is true for Gaussian models. But, if $|\gamma|$ exceeds the standard bound a model will generate data sets with either chaotic behavior relative to model parameters or degenerate data with constant value across the entire spatial domain. Illustrations of this latter are included with several of the examples to come. The importance of standard bounds in interpretation of S-values is that if a given S-value far exceeds a corresponding standard bound, this indicates that the model under contemplation will fail to provide an adequate representation of the data from which it was computed. This aspect of S-value interpretation will be used repeatedly in exploratory examination of the applications of Section 6. Finally, a global measure of the strength of statistical dependence required to capture the spatial structure in a given data set is available as $\gamma/\gamma_{sb}$, which has range $(-1, 1)$ as long as $|\gamma| < \gamma_{sb}$.

# 3 Numerical Investigations of S-value Behavior

In this section we will consider models with binary, Winsorized Poisson, and Gaussian conditionals under $\kappa_i = \kappa$ for all $i$ in the natural parameter functions (2), (3) and (4). The S-value of expression (10) for these models is in the form of a crude estimate of the dependence parameter $\gamma$, although we do not suggest its use as an estimator other than perhaps to obtain starting values for a statistical estimation algorithm. Nevertheless, this does suggest that the statistical behavior of the S-value can be examined through the use of quantities normally associated with the assessment of statistical estimators such as bias and mean squared error. Theoretical derivation of such quantities are not readily available for the S-value, but Monte Carlo assessment is possible.

The exact behavior of the S-value relative to a model of given form depends on many factors such as lattice size, values of the parameters $\kappa$ and $\gamma$, amount of border information excluded, number of bins used (if this is appropriate) and, in some cases, whether or not a restriction is imposed that binned values used to compute $C_\ell$; $\ell = 1, \ldots q$ be used only if the bin sizes $|H_\ell|$ exceed a specified value. Presentation of results concerning all of these factors is beyond the scope of this article, but we give evidence in this section that the S-value is a statistically stable quantity for a number of models.

Simulations were conducted on a $30 \times 30$ lattice, using models with natural parameter functions given by expressions (2), (3), and (4), with $\kappa_i = \kappa$; $i = 1, \ldots, n$ in each case. For each simulated data set, border values of one row and column were used as conditioning values only, that is, were not included in $i = 1, \ldots, n$ but were included in the sets $\boldsymbol{y}(N_i)$; this resulted in $n = 784$ for each data set. A total of $5,000$ data sets were generated for a chosen $\kappa$ and three values of $\gamma$ selected to represent relatively weak, moderate, and strong statistical dependence, for each of the three model types. The simulation sizes of 5000 were sufficient to produce 95% Monte Carlo intervals for the expected S-values with widths less than 5% of the actual $\gamma$ in all cases except a Winsorized Poisson model with weak dependence $\gamma/\gamma_{sb} = 0.1$ (here, $\gamma = 0.0092$, and the interval width was about 7% of this value). Data sets were simulated from a Gibbs algorithm and, in each case, a burn-in of 1000 iterations was used. Every fifth data set after that was collected for use, which was sufficient to eliminate auto-correlation between S-values produced from successive data sets in all situations examined.

For Gaussian models we selected $\kappa = 10$ and conditional variance $\sigma^2 = 1$. For binary models we took $\kappa = 0.5$, and for Winsorized Poisson models we used $\kappa = 5$ and Winsorization value $R = 20$, which is sufficient to produce nearly Poisson behavior in the random field (Kaiser and Cressie, 1997). Values of $\gamma$ used in the simulations correspond to 10%, 50%, and 90% of the standard bounds for $\gamma$ under the various models. Specifically, with any constant $\kappa$ the standard bound for Gaussian models is $1/\sigma^2$ (which is 1 here), so values of $\gamma$ were 0.1, 0.5, and 0.9. With $\kappa = 0.5$ the standard bound for binary models is 4.0, so values of $\gamma$ used were 0.4, 2.0, and 3.6. With $\kappa = 5$ and $R = 20$ the standard bound for Winsorized Poisson models is 0.0924 and we used $\gamma$ with values 0.0092, 0.0462 and 0.0832.

Plots of mean S-value minus $\gamma$, or bias in the sense of estimators, against Monte Carlo simulation size are available for models with Gaussian, binary and Winsorized Poisson conditionals as Web Figure 1 through Web Figure 3. In each case, these plots indicate that the S-value exhibits regular statistical behavior, that is, converges in mean. It is not unbiased for $\gamma$ but, again, we are not proposing its use as such. Monte Carlo estimates of $E\{\text{S-value}\}/\gamma_{sb}$ based on 5000 simulated data sets are presented in Table 1, along with bias, variance, and total squared error (mean squared error). Note from this table that the bias in S-values, if they were to be considered as estimators, is dominated by variance in the assessment of total error.

As mentioned previously, one procedure used to assess the value of including a statistical dependence component in Markov random field models has been a comparison of reduction in prediction mean squared error of a model with dependence from the corresponding model with no dependence, or the so-called independence model. For the models considered here the minimum mean squared error predictors are conditional expectations $E\{Y(\boldsymbol{s}_i)|\boldsymbol{y}(N_i)\} = \tau(A_i)$, so that reduction in prediction mean squared error for a given data set is,

$$R(\gamma) = \frac{1}{n}\sum_{i=1}^{n}\{y(\boldsymbol{s}_i) - \tau(A_i)\}^2 - \frac{1}{n}\sum_{i=1}^{n}\{y(\boldsymbol{s}_i) - \kappa\}^2 . \tag{11}$$

Note that this mean squared error is typically not that for a true prediction, since values $y(\boldsymbol{s}_i)$; $i = 1, \ldots, n$ are observed and used for both estimation and assessment; the quantity in (11) can be considered a fitted mean square, but the term prediction mean squared error has been used in the literature. To determine whether the proposed S-value is related to $R(\gamma)$ we computed (11) for the simulated data sets using the actual parameter values of $\kappa$ and $\gamma$; in practice one would use estimated values of these parameters. Table 1 also presents correlations between S-values and $R(\gamma)$ for the cases considered, and it can be seen that there is a high degree of association between these measures.

The S-value is intended as an exploratory tool for use with individual data sets. That it appears to reflect the magnitude of the model dependence parameter for models that are "well-behaved" is a crucial property in giving assurance that the results in any particular case are meaningful; the standard bounds on values of $\gamma$ are what ensure well-behaved models. But, viewed as a data-generating mechanism, a model may produce individual data sets that do not exhibit the typical or average behavior represented by that model. For Markov random field models such non-typical data sets are those that show spatial structure much weaker or much stronger than would be associated with a given dependence parameter. This latter occurrence is of particular interest. In essence, and as will be illustrated in the next section, there are situations in which an S-value implies greater spatial structure than could reasonably be accommodated by a model with a specified statistical dependence structure. This occurs when the S-value exceeds the standard bound for $\gamma$. It should be kept in mind that, in general, standard bounds for $\gamma$ are not sharp, nor is the S-value an inferential quantity. Thus, in interpreting S-values that exceed a given standard bound on $\gamma$ it would be helpful to have some additional calibration to assist in judging the degree of difference. While far from resolving this issue, a bit of preliminary insight is available from the simulations reported in this section. Empirical distributions of S-values from all 9 cases considered were fairly symmetric, but it is clear that data sets can be easily generated from these models for which the S-value exceeds the standard bound, even if the actual value of $\gamma$ does not. The proportion of data sets producing S-values that were larger than the standard bounds of the corresponding dependence parameter $\gamma$ were 0.1828 for the Winsorized Poisson model, 0.2356 for the binary model, and only 3/5000 for the Gaussian model; recall that the standard bound for the Gaussian model *is* sharp, corresponding exactly to what is needed for a positive definite covariance matrix. For Winsorized Poisson and binary models, the proportion of data sets that result in S-values exceeding the standard bound by percentages from 5% to 40% are presented in Table 2.

Although data sets on which Table 2 is based were simulated from models with dependence parameters at 90% of the standard bounds rather than at the bounds themselves, it can be concluded that data sets with S-values more than 1.2 or 1.25 times the standard bound for $\gamma$ should cause suspicion that more is necessary to adequately represent the data than the model for which the S-value was computed.

# 4 Models with Non-Constant Parameters

The development of the S-value in Section 2 and the basic assessment of its statistical properties in Section 3 considered only the simplest models with natural parameter functions as in expressions (1) but with $\kappa_i = \kappa$ for $i = 1, \ldots, n$. This development can be extended to more complex models in a reasonably straightforward manner. In this section we give two such extensions, one for models with several dependence parameters and the other for models with $\kappa_i$ that vary across locations $\boldsymbol{s}_i; i = 1, \ldots, n$.

## 4.1 Several Dependence Parameters

In many applied problems we want to partition the full neighborhoods into groups of locations such that $N_i = N_i^1 \bigcup N_i^2 \bigcup \ldots \bigcup N_i^G; \; i = 1, \ldots, n$. A common example is partitioning of a four-nearest neighborhood structure into two groups of locations representing horizontal and vertical neighbors to account for directional dependencies. Specifically, if we have $\boldsymbol{s}_i \equiv (u_i, v_i)$ for horizontal coordinate $u_i$ and vertical coordinate $v_i$, we might define $N_i^1 = \{\boldsymbol{s}_j : u_j = u_i \pm 1,\, v_j = v_i\}$ as the horizontal neighborhood and $N_i^2 = \{\boldsymbol{s}_j : u_j = u_i,\, v_j = v_i \pm 1\}$ as the vertical neighborhood. One way to extend the parameterization of (1) to these situations, maintaining $\kappa_i = \kappa;\ i = 1, \ldots, n$ is as,

$$A_i(\boldsymbol{y}(N_i)) = \tau^{-1}(\kappa) + \sum_{g=1}^{G} \gamma_g \frac{1}{m_g} \sum_{\boldsymbol{s}_j \in N_i^g} \{y(\boldsymbol{s}_j) - \kappa\}, \qquad (12)$$

where $m_g;\ g = 1, \ldots, G$ are the sizes of the neighborhood groups so that the total neighborhood size is $m = \sum_g m_g$. Note that some care is needed in formulating models with natural parameter functions as in (12) to ensure that the necessary symmetries among the $\gamma_g$ are satisfied. See Kaiser and Cressie (2000) for necessary conditions for a joint to be identified through use of what is called the negpotential function, and Arnold, Castillo and Sarabia (1992) for more general conditions necessary for the existence of a joint distribution.

A simple and effective way to extend the S-value to model with parameterizations as in (12) is to apply all of the expressions in Section 2 to each neighborhood group individually, ignoring all other groups. That is, consider $G$ sub-models each of the form (12) with only one value of the group index $g$. Then a collection of $G$ S-values results from applying (10) to each sub-model separately, with the appropriate $N_i^g$ and $m_g$ in place of $N_i$ and $m$. Although the full model of expression (12) is not the simple sum of the sub-models, because $\kappa$ appears in the same way in each, the concept of dependence represented by S-values remains unchanged and group S-values computed from the sub-models are interpreted in exactly the same manner as the basic S-values of Section 2, only with respect to the type of dependencies embodied in the group structure (e.g., directional dependencies). Specifically, the magnitude of S-values (and values of the $\gamma_g$) remain interpretable relative to the standard bound available for a given model and $\kappa$. The one additional complication that arises is a need for not only $\gamma_g < \gamma_{sb}$; $g = 1, \ldots, G$ but also the restriction that $\sum_g \gamma_g < \gamma_{sb}$ (see Kaiser, 2007 for details). The use of S-values in detecting directional dependencies will be illustrated in Section 5.2.

## 4.2 Spatial Trend and Covariates

A major advantage of centered parameterizations as in (1) is that, in conjunction with the use of standard bounds for $\gamma$, covariates can be used to further model $\kappa_i$; $i = 1, \ldots, n$ such that the covariate information affects marginal mean structure. If these covariates represent a regular pattern of spatial location this translates into spatial trend. Although several options are available for extending the basic development of the S-value to deal with these situations, we will present only the one we have found the most useful. Suppose that, under the parameterization of expression (1) and the imposition of standard bounds on allowable values of $\gamma$, we have further modeled the large-scale parameters, for $i = 1, \ldots, n$, as $\kappa_i = h(\boldsymbol{x}(\boldsymbol{s}_i), \boldsymbol{\beta})$, for a specified function $h(\cdot)$ and where $\boldsymbol{x}(\boldsymbol{s}_i)$; $i = 1, \ldots, n$ are known covariate values (which might consist of spatial locations $\boldsymbol{s}_i$), and $\boldsymbol{\beta}$ is a vector of unknown parameters. Let $\{\tilde{\kappa}_i : i = 1, \ldots, n\}$ denote preliminary estimates of the $\kappa_i$. Such estimates might be produced through maximum likelihood estimation of typical generalized linear models, ordinary least squares estimation of polynomial regression, or a completely data-driven procedure such as median polish. Note that this latter would constitute an analogous situation with median polish kriging on continuous-index random fields as advocated by Cressie (1993, section 3.5) for some situations involving spatial trend.

In defining the S-value we assumed that neighborhood averages $w(\boldsymbol{s}_i)$ could assume values only in a finite set $\{h_1, \ldots, h_q\}$, and this was either true by definition of the model (e.g., binary, binomial) or by construction of bins (e.g., Gaussian, Poisson). A binning procedure can be applied to the average neighborhood deviations $w^d(\boldsymbol{s}_i) = (1/m)\sum_{\boldsymbol{s}_j \in N_i}\{y(\boldsymbol{s}_j) - \tilde{\kappa}_i\}$ to produce a set of bin midpoints in the same manner described in Section 2.3 for the original $w(\boldsymbol{s}_i)$. Call these bin midoints $h(\ell^d)$; $\ell^d = 1, \ldots, q^d$. This process is repeated for the preliminary estimates $\tilde{\kappa}_i$; $i = 1, \ldots, n$ to arrive at a set of bin midpoints $h(\ell^\kappa)$; $\ell^\kappa = 1, \ldots, q^\kappa$. We then consider the cross-classification of the two sets of bins and create sets $H(\ell^d,\, \ell^\kappa)$, analogous to the $H_\ell$ of Section 2.2, that contain locations $\boldsymbol{s}_i$ for which $w^d(\boldsymbol{s}_i)$ and $\tilde{\kappa}_i$ fall into the cross-classified bins. Some of these sets may be empty or may contain only a small number of values, so we retain only those that have greater than a specified number of observed values. The component quantities $D$ and $r$ of the S-value are then computed using the sets of cross-classified bins. Denote these values as $D(h(\ell^d),\, h(\ell^\kappa))$ and $r(h(\ell^d),\, h(\ell^\kappa))$ to distinguish them from the forms defined explicitly in (7) and (9). The S-value may then be computed exactly as in expression (10) with $D(h(\ell^d),\, h(\ell^\kappa))$ replacing $D(h_\ell, \tilde{\kappa})$ and $r(h(\ell^d),\, h(\ell^\kappa))$ replacing $r(C_\ell, \tilde{\kappa})$. Notational details for creating and cross-classifying this double binning procedure, as well as explicit forms for all quantities involved in computing the S-value for this situation, are presented in Web Appendix A.

# 5 Uses of the S-value

The material of Section 3 indicates that the S-value, despite being a highly "constructed" quantity, does posses regular statistical behavior of the type one expects from a meaningful statistic. As previously indicated, however, the essential value of this diagnostic lies in what it can indicate about statistical dependence in particular data sets, and in this section we present a number of simulated examples designed to illustrate some of the possibilities. It is worth emphasizing that the S-value provides an indication of whether a given data set exhibits spatial structure that is in concert with a proposed model. In other words, the S-value is not intended to detect spatial structure as opposed to the absence of spatial structure. It is intended to provide an indication of whether the structure that might be present in a data set is amenable to modeling through the use of a Markov random field model of specified form. All of the examples of this section were simulated from models based on Winsorized Poisson conditional distributions, but analogous examples with the same behaviors could be produced from any of the three distributional forms that we have considered. The neighborhood structure was set to that of four-nearest neighbors in all of the examples presented, and fitting of models was accomplished through the use of the pseudo-likelihood method of Besag (1974). Because our primary concern in this section is the reflection of data structures by S-values rather than formal inference, interval estimates were produced from the diagonal elements of the inverse hessian based on the log pseudo-likelihood. This should provide a reasonable first approximation, although in an actual application one would want to consider the computation of inferential quantities more carefully.

For a model intended to represent constant mean ($\kappa_i = \kappa; i = 1, \ldots, n$) a preliminary estimate of $\kappa$ is available as the sample mean of all observations. This can be used as a guide in computing a preliminary estimate of the standard bound $\gamma_{sb}$. If a computed S-value is less than this standard bound, then a preliminary estimate of strength of dependence is given by the ratio $S/\gamma_{sb}$. If a computed S-value exceeds the preliminary standard bound, this indicates that one may wish to investigate alternative model structures, such as incorporation of directional dependence or non-constant mean. As demonstrated in Table 2, one should resist using a standard bound, particularly in preliminary form, as an absolute boundary for allowable values. As also demonstrated in Table 2, however, an S-value that far exceeds even a preliminary standard bound should cause one to be quite skeptical about the adequacy of the model under consideration for description of the data. These simple guidelines will play a major role in interpretation of S-values in several of the examples to follow.

## 5.1 Detecting Strength of Dependence

Our first example involves two data sets, both simulated on a $30 \times 30$ lattice from a Winsorized Poisson model having $R = 20$, $\kappa = 5$ and $\gamma = 0.0462$, which is the moderate dependence setting from the simulations of Section 3 ($\gamma/\gamma_{sb} = 0.50$). Plots of the values of $r(C_\ell, \tilde{\kappa}_\ell)$ from expression (9) against values of $D(h_\ell, \tilde{\kappa})$ from expression (7) are presented in the panels of Figure 1, along with the line resulting from the S-value computed from expression (10). Calculated S-values are 0.0788 for the upper data set and 0.0326 for the lower data set in this figure. One would conclude from Figure 1 and these values that the data set corresponding to the upper plot exhibits stronger spatial structure than the data set corresponding to the lower plot. Mean values for the data corresponding to the upper and lower plots of Figure 1 were 4.921 and 5.077, respectively. We use here the standard bound for $\kappa = 5$ and $R = 20$ which is 0.0924, as before. This results in preliminary estimates of the strength of statistical dependence of $0.0788/0.0924 = 0.853$ for the upper case and $0.0326/0.0924 = 0.353$ for the lower case.

Fitting a Winsorized Poisson model with constant parameters $\kappa$ and $\gamma$ to each of these data sets resulted in estimates and 90% intervals of $\hat{\kappa} = 4.854$ (4.637, 5.071) and $\hat{\gamma} = 0.0827$ (0.0578, 0.1075) for data corresponding to the upper plot, and $\hat{\kappa} = 5.065$ (4.906, 5.223) and $\hat{\gamma} = 0.0326$ (0.0102, 0.0551) for the data corresponding to the lower plot of Figure 1. Using estimated values for $\gamma$ and the standard bound of 0.0924, the corresponding estimates of the strength of dependence are $0.0827/0.0924 = 0.895$ and $0.0326/0.0924 = 0.353$, this latter being the same as our preliminary estimate based on S-values and overall sample mean. All of these results indicate that the situation represented by the plot in the upper portion of Figure 1 is one of stronger spatial structure than is that represented by the plot in the lower portion of the figure.

The primary point of this simple example has been to demonstrate that the

S-value reflects the strength of statistical dependence that would be needed to represent the spatial structure present in given data sets. This may differ even among data sets generated from the same model. In addition, the S-values computed for these two data sets were well within a reasonable range of values for a single dependence parameter, indicating that a simple model structure with constant mean and uni-directional dependence should be appropriate (something we already knew based on the model used for simulation).

## 5.2 Detecting Directional Dependence

A data set is presented in the upper left panel of Figure 2 that was simulated on a $30 \times 30$ lattice from a Winsorized Poisson model having natural parameter function as in expression (12) with $\kappa = 5$, $G = 2$, $\gamma_1 = 0.07$, $\gamma_2 = 0.001$, and $m_1 = m_2 = 2$. The neighborhood groups were defined to be directional, with $N_i^1 = \{\boldsymbol{s}_j : u_j = u_i \pm 1, v_j = v_i\}$ representing a horizontal neighborhood and $N_i^2 = \{\boldsymbol{s}_j : u_j = u_i, v_j = v_i \pm 1\}$ representing a vertical neighborhood. The data set of Figure 2 should contain directional dependence, stronger in the horizontal direction and weaker (in fact nearly absent) in the vertical direction, although this is not clearly evident from the image plot of the data. The upper right panel of Figure 2 shows $r(C_\ell, \tilde{\kappa}_\ell)$ versus $D(h_\ell, \tilde{\kappa})$ for a model with a single dependence parameter, and a solid line indicating the calculated S-value of 0.0875. As the overall data mean is 5.144, this suggests strength of dependence 0.9470, which is quite high; we again used the preliminary standard bound of 0.0924. The lower left panel of Figure 2 shows the plot of these values calculated using only neighbors in the horizontal neighborhood $N_i^1$ and has S-value 0.0700 or strength of dependence 0.7576, and the lower right panel plots the corresponding values for the vertical neighborhood $N_i^2$ which result in S-value 0.0008 and strength of dependence 0.0086. One implication of these results is that it should be entirely possible to fit a model with a single dependence parameter $\gamma$ to these data, as evidenced by the unidirectional S-value. Doing so results in estimated parameter values and 90% intervals of $\hat{\kappa} = 5.071$ (4.826, 5.316), $\hat{\gamma} = 0.0899$ (0.0693, 0.1105), and estimated strength of dependence 0.9729. But the S-value examination of the data also indicates the difference in strength of dependence between horizontal and vertical directions and that fitting a model with two directional dependence parameters could be potentially valuable. Doing this results in estimated parameters and intervals of $\hat{\kappa} = 5.018$ (4.782, 5.254), $\hat{\gamma}_1 = 0.0854$ (0.0710, 0.0997), and $\hat{\gamma}_2 = 0.0010$ ($-0.0142$, 0.0163), with corresponding estimates of the differential dependence of 0.9242 and 0.108, respectively. Particularly given our knowledge of the true data generating mechanism, the directional model seems a better representation of the data. Were this an actual data set, we might reasonably conclude that not only is dependence directional, but only the horizontal direction requires modeling.

## 5.3 Detecting Spatial Trend

Figure 3 presents a data set simulated from a Winsorized Poisson model that contains spatial trend in the large-scale model component. The natural parameter function for this model was that of expression (1) with $\tau^{-1}(\kappa_i) = \log(\kappa_i)$, $\kappa_i = 0.15(u_i + v_i)$; $i = 1, \ldots, n$, and $\gamma = 0.05$. The upper left panel of Figure 3 contains an image plot of the data and, in this case, a visual indication of large-scale structure might seem apparent. However, as noted in the Introduction, spatial structure, even if it may appear to be in the form of a trend, can sometimes be modeled through dependence only. We might wonder if this is the case for these data, particularly if a scientific explanation for the apparent trend is not readily available (under whatever hypothetical problem one wishes to imagine for these data). The overall data average for the values in the upper left of Figure 3 is 4.681 giving a preliminary standard bound of 0.0948 (or 0.0924 as previously if we just use 5 for the calculation as done in previous examples). The upper right panel of Figure 3 presents the plot of $r(C_\ell, \tilde{\kappa}_\ell)$ versus $D(h_\ell, \tilde{\kappa})$ assuming a model with constant $\kappa$, and the resultant S-value is 0.1744, which is over 1.8 times the preliminary standard bound. This seems quite a large value. A model with $\kappa = 4.681$ and $\gamma = 0.1744$ would not lead to data sets that have average values anywhere near 4.681. In fact, the minimum, average, and maximum data means in 2000 data sets simulated with these parameter values were 19.938, 19.984 and 20.000, respectively; recall the Winsorization value of $R = 20$. Thus, the S-value has provided a clear indication that a model with constant mean and a single dependence parameter is not a tenable choice to describe these data.

The lower left panel of Figure 3 presents values of $r(h(\ell^d), h(\ell^\kappa))$ described in Section 4.2 plotted against values of $D(h(\ell^d), h(\ell^\kappa))$ when initial estimates of $\tilde{\kappa}_i$; $i = 1, \ldots, n$ are produced by a median polish algorithm. Here, the resultant S-value is 0.0685. The lower right panel of Figure 3 presents the corresponding plot when initial estimates of the $\tilde{\kappa}_i$ are produced from an ordinary least squares fit of $\kappa_i = \beta_0 + \beta_1 u_i + \beta_2 v_i$, which results in $\hat{\beta}_0 = -0.2039$, $\hat{\beta}_1 = 0.1577$ and $\hat{\beta}_2 = 0.1574$. In this case the S-value is 0.0450.

Overall, use of the S-value in this example has provided solid evidence that a model with constant mean would not be adequate to represent these data, while a model with some form of spatial trend in the large-scale model structure likely would be appropriate. Estimated parameter values and 90% intervals for the model using the linear regression for $\kappa_i$; $i = 1, \ldots, n$ are $\hat{\beta}_0 = -0.049$ $(-0.349, 0.250)$, $\hat{\beta}_1 = 0.148$ $(0.129, 0.164)$, $\hat{\beta}_2 = 0.152$ $(0.134, 0.171)$, and $\hat{\gamma} = 0.0500$ $(0.0273, 0.0728)$. These estimates agree well with the model having $\beta_0 = 0$, $\beta_1 = \beta_2 = 0.15$ and $\gamma = 0.05$. The representation of strength of statistical dependence as an estimate of $\gamma$ divided by its standard bound depends on the value of $\kappa$ used to calculate the standard bound (except for Gaussian models) which complicates estimating this strength of dependence in cases with non-constant $\kappa_i$, but we might use the largest value of $\kappa_i$ as a conservative estimate. Under the fitted model the largest value of $\hat{\kappa}_i$ is 8.591, and our final estimate of the standard bound for $\gamma$ would be $\{\log(20) - \log(8.591)\}/(20 - 8.591) = 0.0728$ with a corresponding estimate of strength of dependence $0.0500/0.0728 = 0.6871$.

# 6 Exploratory Analysis in Applications

In this section we re-examine several published applications of Markov random field models to spatial problems. These applications involve situations in which authors have fit a range of models to determine what might be an appropriate model structure, including both large-scale and small-scale model components. Our objective is not to contrast another full analysis of these problems with previously published results, but rather to demonstrate what might be discovered in these problems through application of the exploratory S-value.

## 6.1 Drumlins In Ireland

Griffith (2006) overlaid $11 \times 11$ grids on three $64\text{km}^2$ portions of County Down in Northern Ireland, for which Hill (1973) had geo-referenced locations of individual landforms called drumlins, which are ridges or oval-shaped hills formed by glacial movements. The number of drumlins in each grid cell was tabulated resulting in a regular lattice of count data for each of the three regions. We will call these Region 1, Region 2, and Region 3. According to Griffith (2006) what we are calling Region 1 corresponds to a plot in the upper Ards peninsula, Region 2 to a plot west of Strangford Lough, and Region 3 to a plot east of Slieve Croob (see Hill, 1973 for maps of these areas). The correspondence of the three regions with physical locations will be important in our exploratory analysis.

Hill (1973) reports that the larger County Down area contains two major glacial till sheets, one deposited by a North Channel ice sheet moving from the north and northeast to the south and southwest, and a more recent till sheet deposited by an Irish ice sheet moving from northwest to southeast (see Figure 2 in Hill, 1973). Based on morphologies of individual drumlins, Hill (1973) suggests that drumlins in north Ards peninsula (our Region 1) were formed by the North Channel ice sheet, while those in other portions of County Down (our Regions 2 and 3) were more likely the result of the Irish ice sheet. Hill also reports that the relation of drumlins to directions of ice movement "is complicated" (Hill, 1973, p. 229). His analysis suggests that, aside from north Ard penisula, there might be bands of drumlin intensity that are oriented in a northeast to southwest direction, which would be perpendicular to the movement of the Irish ice sheet believed to have formed these drumlins. Our concern here will be to determine what suggestions regarding these issues of spatial structure in drumlin intensities could be gleaned from an exploratory analysis based on S-values.

The data means for the three regions were 1.934 for Region 1, 1.942 for Region 2 and 1.264 for Region 3, and we chose a Winsorization value of $R = 7$, the same value used by Griffith (2006). S-values were computed based on 6 bins which, with 81 interior values on an $11 \times 11$ lattice gives about 13 values per bin; we were reluctant to use more bins with smaller numbers of observations per bin. Preliminary standard bounds were then 0.2539, 0.2535, and 0.2984 for the three regions, respectively. Particularly given the small lattices involved, we took the standard bound to be in the interval (0.25, 0.30) rather than assigning one specific value.

The three lattices for Regions 1 through 3 were oriented with horizontal coordinate in an east-west direction and vertical coordinate in a north-south direction. Given this, we computed S-values for situations in which dependence was taken as unidirectional, in the north-south direction, in the east-west direction, in the northeast-southwest direction, and in the northwest-southeast direction. The results are presented in Table 3, from which it may be seen that the unidirectional S-values for Region 1 and Region 2 are well above even a liberally chosen standard bound of 0.30, suggesting that models with constant mean and one dependence parameter would not be tenable for the data in these regions. Similarly, a model with directional dependencies in the east-west and north-south directions would not be realistic for these regions, as the sum of S-values for those directions also well exceed standard bounds. In contrast, Region 3 appears that it could be modeled with a unidirectional dependence or dependencies in the primary compass directions. To emphasize these conclusions, 2000 data sets were simulated for each region with unidirectional dependence and with directional dependence in the east-west and north-south directions, using dependence parameters as given by the corresponding S-values. The Monte Carlo average of data means for Region 1 were 5.28 and 6.97 for the unidirectional and directional cases, respectively; the actual data mean is 1.934. For Region 2 these values were 6.82 and 5.70 in the same order; the actual data mean is 1.942. In contrast, values for Region 3 were 1.28 and 1.28 with an actual data mean of 1.264. Web Figure 4 presents boxplots of all of the data set means for these 2000 simulated data sets.

Considering dependencies in other directions results in interesting suggestions relative to the analysis of Hill (1973). As indicated in Table 3, Region 1 appears to have fairly strong dependence in the northeast-southwest direction but much less dependence in the northwest-southeast direction. Regions 2 and 3, in contrast, would seem to have dependencies stronger in the northwest-southeast direction, clearly so for Region 3; for Region 2 these directional S-values are more similar. Final conclusions are inappropriate as this is intended to be an exploratory analysis. But, the strong northeast-southwest dependence in Region 1 agrees with the direction of ice flow for the North Channel ice sheet that Hill (1973) suggests produced these drumlins. Similarly, the strong northwest-southeast dependence in Region 3 agrees with the direction of ice flow for the Irish ice sheet deemed primarily responsible for drumlins in the central portion of County Downs. Results for Region 2 are more equivocal, demonstrating moderate dependencies in both the northwest-southeast and northeast-southwest directions, although perhaps a bit stronger along the northwest-southeast gradient. These results agree with those of Hill (1973) in detecting a difference between north Ards peninsula (Region 1) and the rest of County Downs, but may be at odds with the detected spatial "bands" of drumlins running northeast-southwest through most of the area. Our results suggest that spatial structure in these drumlin fields, as modeled by statistical dependence, might be parallel to the direction of ice flow, not perpendicular as asserted by Hill. Note again, however, that we were working with only a fraction of the entire region considered by Hill.

From a purely statistical viewpoint, the outcome of this exploratory treatment of the Irish drumlin data is that it appears appropriate to model the data from Regions 1 and 2 with a model having dependence structures in the northwest-southeast and northeast-southwest directions, but not with dependence structures in the primary compass directions or with unidirectional dependence. In constrast, the suggestion is that data from Region 3 might be adequately modeled using dependencies that are unidirectional, following the primary compass directions, or following northwest-southeast and northeast-southwest gradients. These suggestions are verified in the boxplots of overall means for data sets simulated from these models and presented in Web Figure 4.

## 6.2 Plant Disease in Agricultural Fields

Graham (1996) and later Gumpertz *et al.* (1997) examined the prevalence of a plant disease in fields of green peppers through the use of auto-logistic models, or Markov random field models with binary conditionals. Graham (1996) considered a subset of data used in the latter analysis. His conclusions included that there appeared to be a difference in dependence between directions that ran along cultivation rows and across cultivation rows. Gumpertz *et al.* (1997) concluded that both covariates and statistical dependence terms were needed to model the patterns of disease incidence they observed. Here, we re-examine the data of Graham (1996) through the use of exploratory S-values.

The overall sample mean for the data of Graham (1996) is 0.36 and the basic S-value computed for a model with constant mean and a single unidirectional dependence parameter with four nearest neighbors is 4.6, well above the standard bound of 4.08 for this mean. This suggests that fitting a simple model with unidirectional dependence is most likely not a productive pursuit for these data. S-values for a model with two dependence parameters, one in the horizontal direction and one in the vertical direction similar to the example of Section 5.2, produces $S_u = 2.5$ and $S_v = 3.6$. While both of these values are below the standard bound their sum is 6.1, indicating again that this model is not viable for description of the data. Using a fitted logistic regression to produce preliminary estimates of $\{\kappa_i : i = 1, \ldots, n\}$, a unidirectional S-value as described in Section 4.2 results in a value of 2.13. The implications of this simple exploratory examination of the data strongly suggests that neither a model with unidirectional dependence and constant mean, nor a model with directional dependence (along and across rows in the field) and a constant mean are likely to result in adequate descriptions of the data. To further emphasize these conclusions, 2000 data sets were again simulated from the three possiblities, constant mean with either one or two dependence parameters, and unidirectional dependence but with $\log(\kappa_i) - \log(1 - \kappa_i) = \beta_0 + \beta_1 u_i + \beta_2 v_i$. The results regarding sample means (proportions) are summarized in the boxplots of Web Figure 5. Neither of the models with constant mean can produce data sets with realized proportions near the actual value in the data, while the model that explicity accounts for spatial trend is able to do so. This supports the conclusion of Gumpertz *et al.* (1997) that it is important to account for covariate information in this problem.

## 7 Concluding Remarks

We have introduced an exploratory quantity we call the S-value that is useful in determining whether spatial structure exhibited by data is amenable to modeling through the use of one-parameter exponential family Markov random field models. We have demonstrated that the S-value possesses regular statistical behavior as would be expected from any meaningful statistic, and have shown its potential uses through a number of simulated scenarios. Finally, we have also demonstrated its usefulness in guiding the modeling process through exploratory consideration of several previously published applications.

Although the S-value appears to be a quite useful quantity, any number of questions can be posed regarding the details of its implementation and possible limitations. Our simulated examples were all produced for lattices of moderate size, $30 \times 30$. The applications of Section 6 involved smaller lattices, and the S-value seems to have performed admirably in these two examples. But the overall effect on S-values of lattice size and procedures for handling border information is not completely understood. For situations that involve binning, trade offs between the number of bins and the numbers of observations within those bins is an issue in need of additional investigation. Robust versions of the S-value might be contemplated by replacing the ordinary least squares value of expression (10) with a robust fit of $D$ to $r$, although preliminary investigations in this direction have not been overly promising. It would also be of great use to extend the S-value concept to models in which a Markov random field is overlaid on the parameters of a conditionally independent data model. While questions such as these remain in need of additional investigation, it seems apparent that the S-value has a great deal of use in indicating model structures that should prove beneficial to consider in the overall analysis of given data sets.

## Supplementary Materials:

Web appendices and figures referenced in Sections 2, 3, 4, and 6 are available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowlegements: The authors thank Daniel Griffith for providing access to the data on Irish drumlins.

## References

Arnold, B.C., Castillo, E. and Sarabia, J. (1992), *Conditionally Specified Distributions.* Lecture Notes in Statistics, Vol. 73, Springer-Verlag, Berlin.

Augustin, N.H., McNicol, J. and Marriott, C.A. (2006), Using the truncated auto-Poisson model for spatially correlated counts of vegetation, *Journal of Agricultural, Biological, and Environmental Statistics* **11**, 1-23.

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

Caragea, P.C. and Kaiser, M.S. (2007), Covariates and time in the autologistic model, Preprint 2006-19, Department of Statistics, Iowa State University, Ames, Iowa.

Cressie, N.A.C. (1993), *Statistics for Spatial Data*, rev. ed., Wiley: New York.

Furukawa, K. (2004), Development of Markov Random Field Models Based on Exponential Family Conditional Distributions, Unpublished PhD dissertation, Iowa State University, Ames, Iowa.

Graham, J.M. (1996), Markov chain Monte Carlo methods for modeling the spatial pattern of disease spread in bell pepper, pp. 91-108 in Proceedings of the

1996 Kansas State University Conference on Applied Statistics in Agriculture, Department of Statistics, Kansas State University, Manhattan, Kansas.

Griffith, D.A. and Layne, L.J. (1999), *A Casebook for Spatial Statistical Data Analysis*, Oxford University Press, Oxford.

Griffith, D.A. (2006), Assessing spatial dependence in count data: Winsorized and spatial filter specification alternatives to the auto-Poisson model, *Geographical Analysis* **38**, 160-179.

Gumpertz, M.L., Graham, J.M., Ristaino, J.B. (1997), Autologistic model of spatial pattern of Phytophtora epidemic in bell pepper: effects of soil variables on disease presence, *Journal of Agricultural, Biological and Environmental Statistics* **2**, 131-156.

Haining, R. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press: Cambridge, United Kingdom.

Hill, A.R. (1973), The distribution of drumlins in County Down, Ireland, *Annals of the Association of American Geographers* **63**, 226-240.

Hoeting, J.A., Leecaster, M. and Bowden, D. (2000), An improved model for spatially correlated binary responses, *Journal of Agricultural, Biological, and Environmental Statistics* **5**, 102-114.

Kaiser, M.S. (2007) Statistical dependence in Markov random field models, Preprint 2007-1, Department of Statistics, Iowa State University, Ames, Iowa.

Kaiser, M.S. (2001), Markov random field models. In A.H. El-Shaarawi and W.W. Piegorsch, eds. *Encyclopedia of Environmetrics*, Wiley and Sons, New York.

Kaiser, M.S., Daniels, M.J., Furukawa, K. and Dixon, P. (2002), Analysis of particulate matter air pollution using Markov random field models of spatial dependence, *Environmetrics* **13**, 615-628.

Kaiser, M.S. and Cressie, N. (2000), The construction of multivariate distributions from Markov random fields, *Journal of Multivariate Analysis* **73**, 199-220.

Rue, H. and Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, Chapman & Hall/CRC Press: Boca Raton, Florida.

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

Wikle, C.K., Milliff, R.F., Nychka, D. and Berliner, L.M. (2001), Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds, *Journal of the American Statistical Association* **96**, 382-397.

Wu, H. and Huffer, F.W. (1997), Modeling the distribution of plant species using the autologistic regression model, *Environmental and Ecological Statistics* **4**, 49-64.

Zhu, J., Huang, H-C. and Wu, J. (2005), Modeling spatial-temporal binary data using Markov random field models, *Journal of Agricultural, Biological, and Environmental Statistics* **10**, 212-225.

# Tables and Figures

Table 1: Monte Carlo results for 5000 simulated data sets in each of nine situations involving three distributional forms and three levels of statistical dependence. See text for complete description.

| Model | $\gamma/\gamma_{sb}$ | $E_{MC}$ | $B_{MC}$ | $var_{MC}$ | $mse_{MC}$ | $cor(S, R(\gamma))$ |
|---|---|---|---|---|---|---|
| Gaussian | 0.90 | 0.86 | -0.04 | 0.0023 | 0.0040 | 0.78 |
| | 0.50 | 0.47 | -0.03 | 0.0068 | 0.0076 | 0.96 |
| | 0.10 | 0.09 | -0.01 | 0.0089 | 0.0090 | 0.99 |
| Binary | 0.90 | 0.92 | 0.02 | 0.0146 | 0.0148 | 0.79 |
| | 0.50 | 0.51 | 0.01 | 0.0141 | 0.0141 | 0.87 |
| | 0.10 | 0.10 | 0.00 | 0.0123 | 0.0123 | 0.90 |
| Poisson | 0.90 | 0.83 | -0.07 | 0.0342 | 0.0388 | 0.90 |
| | 0.50 | 0.45 | -0.05 | 0.0419 | 0.0440 | 0.95 |
| | 0.10 | 0.08 | -0.02 | 0.0435 | 0.0439 | 0.97 |

Table 2: Proportion of simulated data sets out of 5000 for which the computed S-value exceeded the standard bound on $\gamma$ by various percentages.

| | Percentage of $\gamma_{sb}$ by which S-value exceeds $\gamma_{sb}$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | 5% | 10% | 15% | 20% | 25% | 30% | 35% | 40% |
| Binary | 0.1326 | 0.0722 | 0.0360 | 0.0132 | 0.0054 | 0.0026 | 0.0010 | 0.0006 |
| Poisson | 0.1176 | 0.0698 | 0.0370 | 0.0178 | 0.0094 | 0.0040 | 0.0020 | 0.0008 |

Table 3: S-values for the data on Irish drumlins.

| | Calculated S-Value | | | | |
|---|---|---|---|---|---|
| Region | Unidirectional | N-S | E-W | NE-SW | NW-SE |
| 1 | 0.3681 | 0.3848 | 0.1848 | 0.2451 | 0.0688 |
| 2 | 0.3763 | 0.1574 | 0.3086 | 0.0948 | 0.1435 |
| 3 | 0.2197 | 0.0104 | 0.1961 | 0.0125 | 0.2015 |

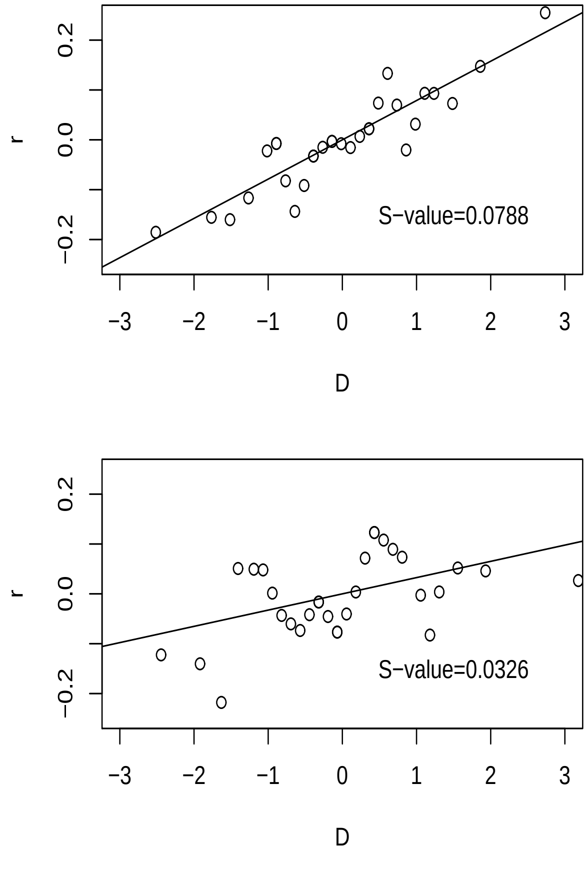


Figure 1: Plots of component quantities $r$ and $D$ for computation of S-values for two data sets from a Winsorized Poisson model with $\kappa = 5$ and $\gamma = 0.462$. Slopes of the lines are S-values for these cases.

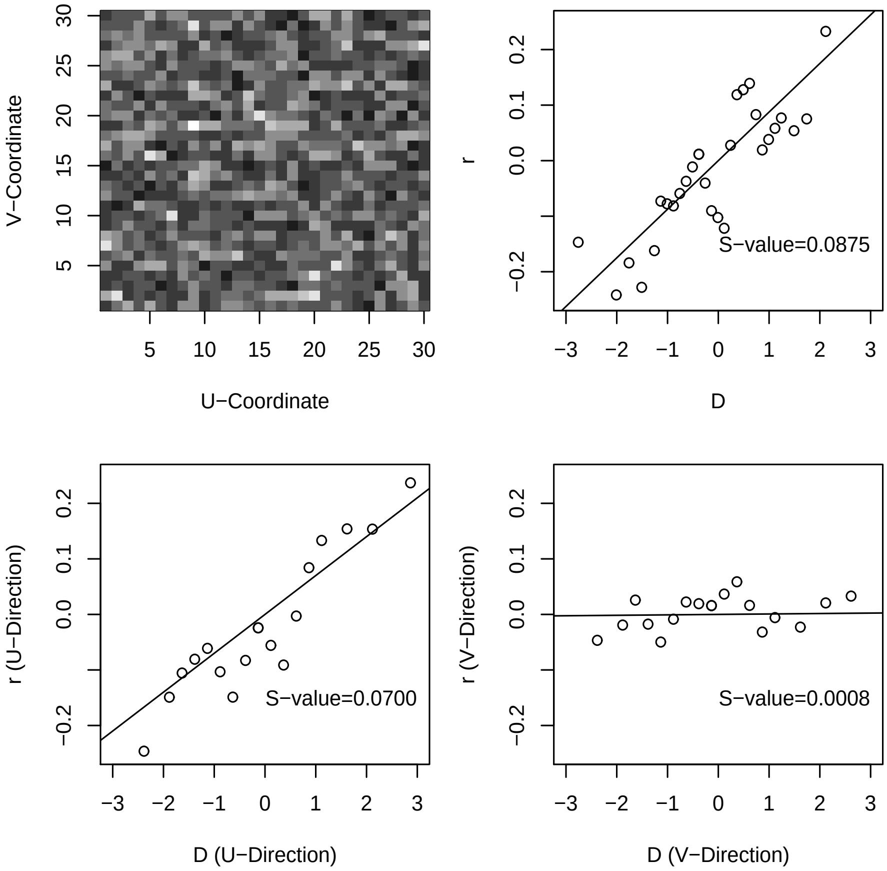


Figure 2: Example of data exhibiting directional dependence. Image plot of data is given in upper left panel, and component quantities of the S-value are shown for unidirectional dependence in the upper right, horizontal dependence in the lower left, and vertical dependence in the lower right.

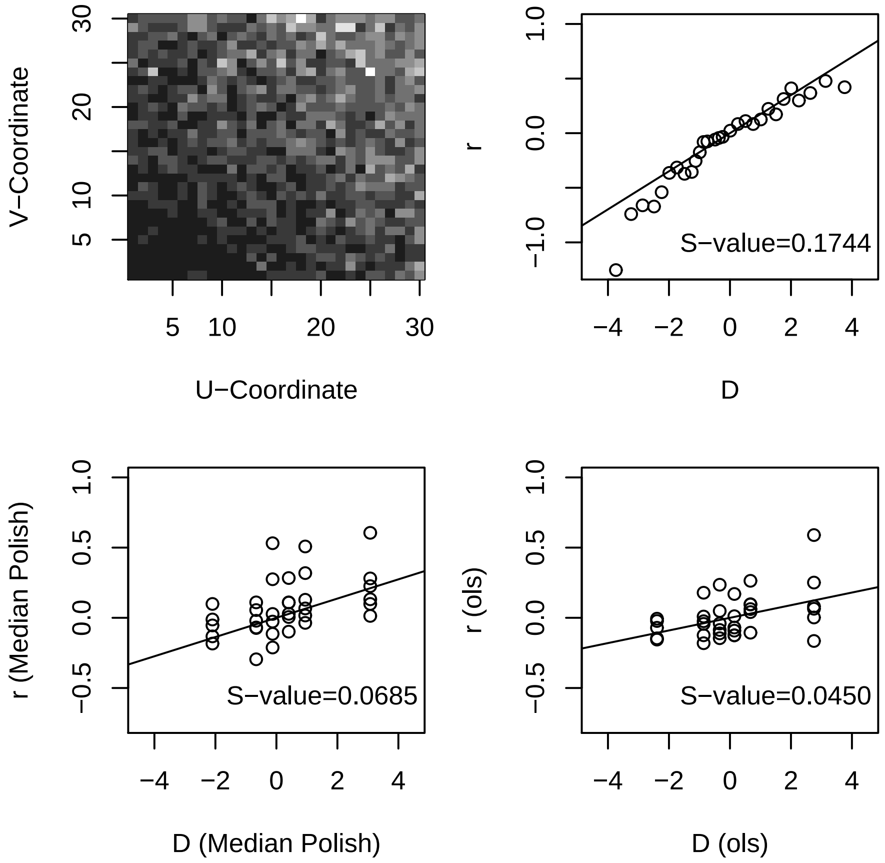


Figure 3: Example of data with spatial trend. Image plot of data is given in upper left panel, component quantities of S-value for model with constant mean in upper right panel, for model using median polish in lower left, and for model with ordinary least squares estimates in lower right.

# Web-Based Supplementary Materials

Exploring Dependence with Data on Spatial Lattices

M.S. Kaiser and P.C. Caragea

## 1 Web Appendix A

### 1.1 Single Binning Procedure of Section 2.2

In Section 2.2 of the article we briefly describe dividing the values $w(\boldsymbol{s}_i)$; $i = 1, \ldots, n$ into bins for the purpose of computing the quantities $h_\ell$; $\ell = 1, \ldots, q$ used in construction of the S-value of expression (10). Notational details of this procedure are presented here.

Given observations $\{y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ and neighborhoods $\{N_i : i = 1, \ldots, n\}$, the values of $h_\ell$; $\ell = 1, \ldots, q$ can be computed by dividing the empirical distribution of the observed neighborhood averages

The values to be binned are neighborhood averages $w(\boldsymbol{s}_i) = (1/m) \sum_{\boldsymbol{s}_j \in N_i} y(\boldsymbol{s}_j)$; $i = 1, \ldots, n$. Let $\{\omega_j : j = 1, \ldots, q+1\}$ denote ordered quantiles of the empirical distribution of these values, with $\omega_1 = \min\{w(\boldsymbol{s}_i)\}$ and $\omega_{q+1} = \max\{w(\boldsymbol{s}_i)\}$. Then, for $\ell = 1, \ldots, q$ let $h_\ell = (\omega_{\ell+1} + \omega_\ell)/2$ and $H_\ell = \{y(\boldsymbol{s}_i) : \omega_\ell \leq w(\boldsymbol{s}_i) \leq \omega_{\ell+1}\}$.

### 1.2 Double Binning Procedure of Section 4.2

Section 4.2 of the manuscript briefly describes a double binning procedure for computation of the quantities $h(\ell^d)$; $\ell^d = 1, \ldots, q^d$ and $h(\ell^\kappa)$; $\ell^\kappa = 1, \ldots, q^\kappa$ which are used to compute an S-value for models with non-constant large scale structure $\kappa_i$; $i = 1, \ldots, n$. Details of that procedure are given here.

Define average neighborhood deviations as $w^d(\boldsymbol{s}_i) = (1/m) \sum_{\boldsymbol{s}_j \in N_i} \{y(\boldsymbol{s}_j) - \tilde{\kappa}_j\}$. Let $\{\omega(\ell^d) : \ell^d = 1, \ldots, q^d + 1\}$ denote ordered quantiles of the empirical distribution of $\{w^d(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ and define $h(\ell^d) = (\omega(\ell^d) + \omega((\ell+1)^d)$; $\ell^d = 1, \ldots, q^d$ as bin midpoints. This process is repeated for the preliminary estimates $\tilde{\kappa}_i$; $i = 1, \ldots, n$ to arrive at a set of quantiles $\{\omega(\ell^\kappa) \,:\, \ell^\kappa = 1, \ldots, q^\kappa + 1\}$ and bin midpoints $h(\ell^\kappa) = (\omega(\ell^\kappa + \omega((\ell+1)^\kappa))$; $\ell^\kappa = 1, \ldots, q^\kappa$.

We then consider the cross-classification of the two sets of bins and create sets analogous to the $H_\ell$ of Section 2.2 as,

$$H(\ell^d, \ell^\kappa) = \{\boldsymbol{s}_i : \omega(\ell^d) \leq w^d(\boldsymbol{s}_i) \leq \omega((\ell+1)^d);\ \omega(\ell^\kappa) \leq \tilde{\kappa}_i \leq \omega((\ell+1)^\kappa)\}, \quad \text{(A.1)}$$

for $1 \leq \ell^d \leq q$ and $1 \leq \ell^\kappa \leq q$. Some of these sets may be empty or may contain only a small number of values, so we retain only those that have greater than a specified number of observed values $s$ as, $H(\boldsymbol{z}_p) \equiv \{H(\ell^d, \ell^\kappa) \,:\, |H(\ell^d, \ell^\kappa)| \geq s\}$ where $\boldsymbol{z} \equiv (\ell^d, \ell^\kappa)$ and we re-index as $p = 1, \ldots, k$. The S-value is then computed by replacing $D(h_\ell, \tilde{\kappa})$; $\ell = 1, \ldots, q$ in expression (7) with

$$D(\boldsymbol{z}_p) = h(\ell^d)\, I\left(\boldsymbol{z}_p = (\ell^d, \cdot)\right);\ p = 1, \ldots, k, \quad \text{(A.2)}$$

where $I(\cdot)$ is the indicator function. Note that this implies potential multiple uses of the specific values $h(\ell^d)$. Similarly, the $C_\ell$; $\ell = 1, \ldots, q$ of expression (8) are replaced with, for $p = 1, \ldots, k$,

$$C(\boldsymbol{z}_p) = \frac{1}{|H(\boldsymbol{z}_p)|} \sum_{\boldsymbol{s}_i \in H(\boldsymbol{z}_p)} Y(\boldsymbol{s}_i), \quad \text{(A.3)}$$

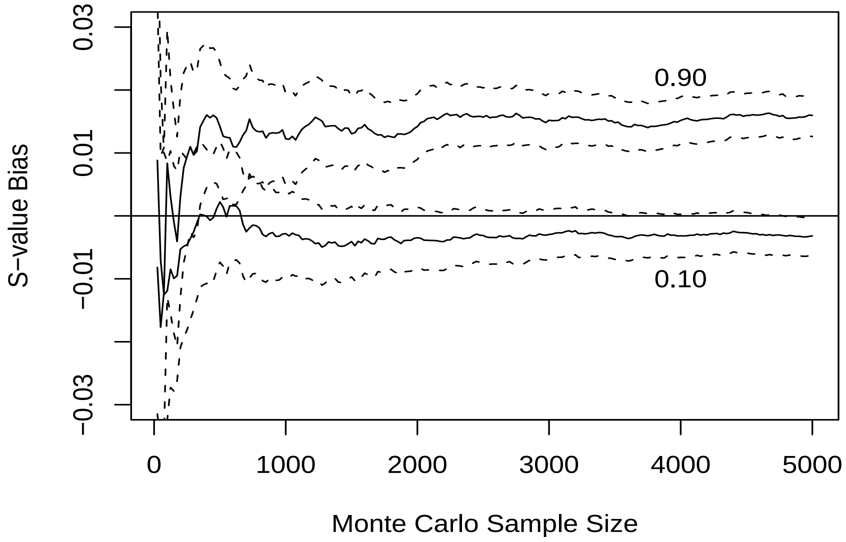
and the $r(C_\ell, \tilde{\kappa})$ of expression (9) are replaced with, for $p = 1, \ldots, k$,

$$r(\boldsymbol{z}_p) = \tau^{-1}(C(\boldsymbol{z}_p)) - \tau^{-1}(h(\ell^\kappa)). \quad \text{(A.4)}$$

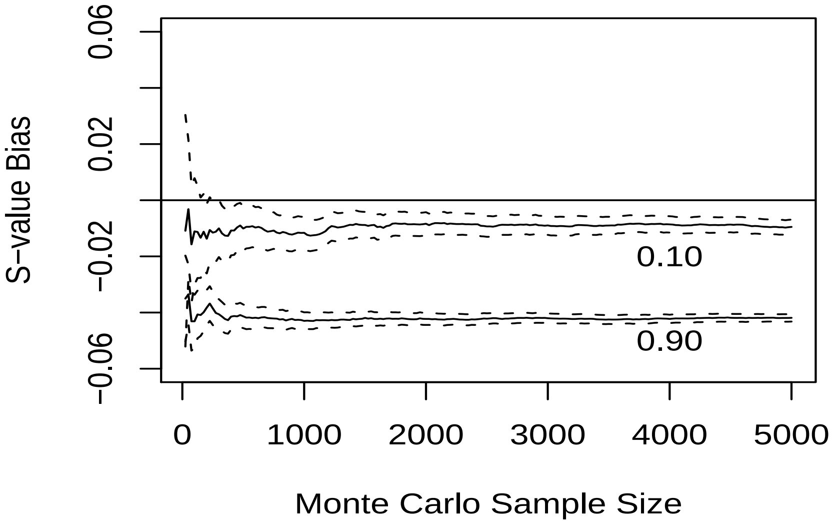
The S-value of expression (10) then becomes,

$$S = \frac{\sum_{p=1}^{k} r(\boldsymbol{z}_p) D(\boldsymbol{z}_p)}{\sum_{p=1}^{k} \{D(\boldsymbol{z}_p)\}^2}. \quad \text{(A.5)}$$

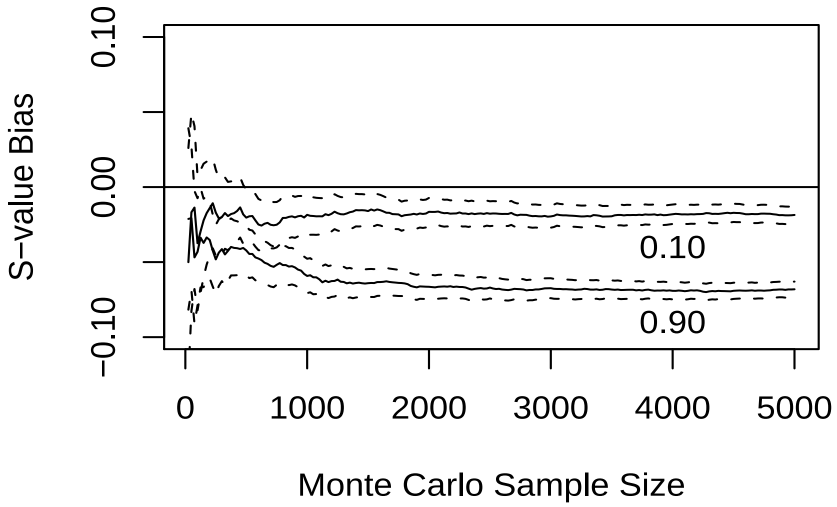
# 2 Web Figures



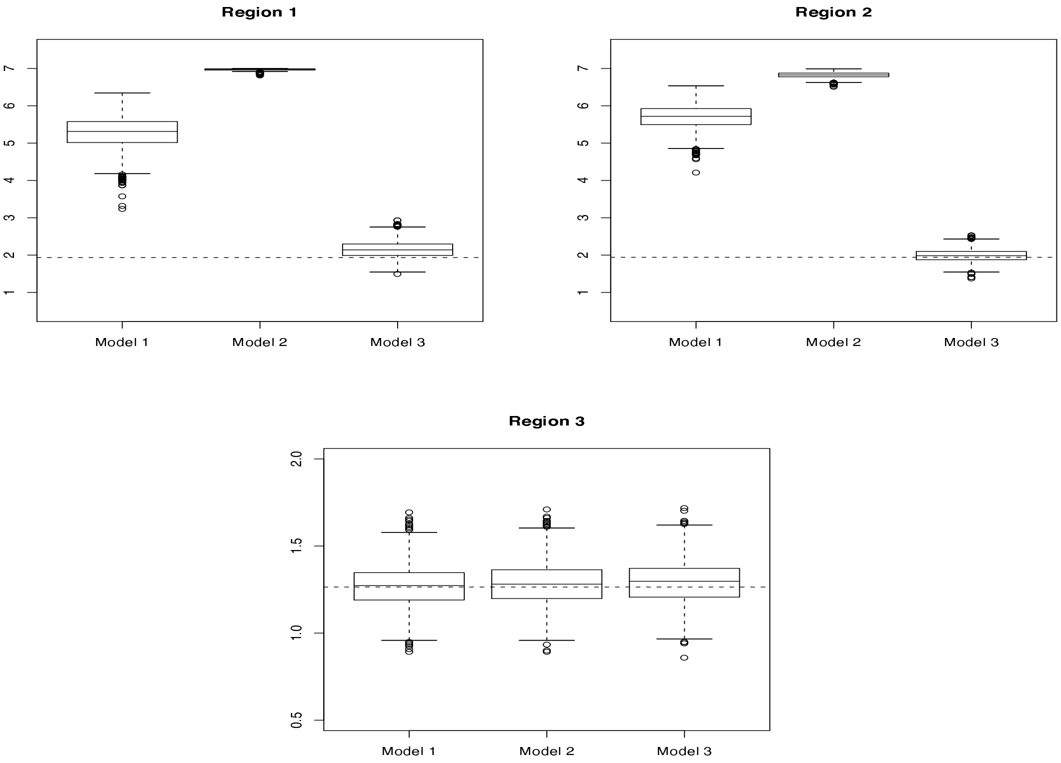
Web Figure 1. Plot of S-value bias over Monte Carlo simulations for a binary model with $\kappa = 0.5$ and values $\gamma/\gamma_{sb} = 0.90$ and 0.10. Solid lines are Monte Carlo expected values, dashed lines are 95% Monte Carlo intervals for the expected values.



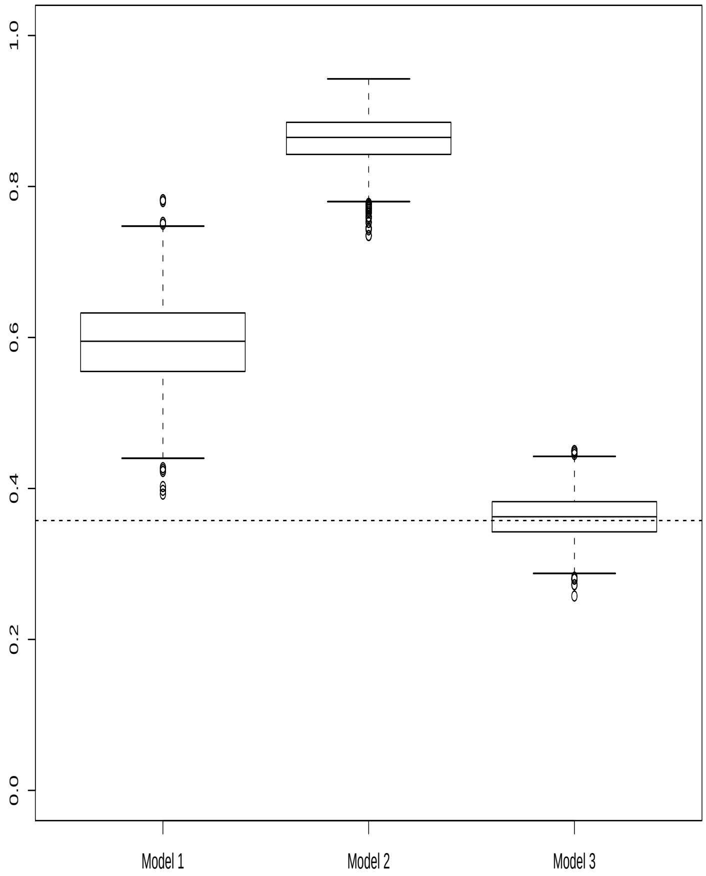
Web Figure 2. Plot of S-value bias over Monte Carlo simulations for a Gaussian model with $\kappa = 10$ and values $\gamma/\gamma_{sb} = 0.90$ and 0.10. Solid lines are Monte Carlo expected values, dashed lines are 95% Monte Carlo intervals for the expected values.



Web Figure 3. Plot of S-value bias over Monte Carlo simulations for a Winsorized Poisson model with $\kappa = 5$ and values $\gamma/\gamma_{sb} = 0.90$ and 0.10. Solid lines are Monte Carlo expected values, dashed lines are 95% Monte Carlo intervals for the expected values.



Web Figure 4. Boxplots of data set average values for simulated data sets corresponding to possible models for the Irish drumlin data. The three sets of side-by-side boxplots correspond to the three regions of data as titled in the figure. In each case, Model 1 corresponds to unidirectional dependence, Model 2 to directional dependence in the north-south and east-west directions, and Model 3 to directional dependence in the northwest-southeast and northeast-southwest directions. Values of dependence parameters used to simulated data are given as S-values in Table 3 of the article. Actual mean values of the observed data were used as values of $\kappa$ in the simulation models, and are represented as horizontal dashed lines in the figure.



Web Figure 5. Boxplots of data set average values for simulated data sets corresponding to possible models for the Bell Pepper data. Model 1 corresponds to constant mean with unidirectional dependence. Model 2 corresponds to constant mean with directional dependence within and across cultivation rows. Model 3 corresponds to spatial trend in the large-scale model component and unidirectional dependence. The actual data mean is shown as the horizontal dashed line.