ISSN 1946-7664. MCFNS 2017
Vol. 9, Issue 2, pp. 17–??
Mathematical and Computational Forestry & Natural-Resource Sciences

AVAILABLE ONLINE AT HTTP://MCFNS.COM

Submitted: Aug. 23, 2016
Accepted: Jul. 11, 2017
Published: Sep. 30, 2017
Last Correction: Sep. 31, 2017

INCORPORATING A LOCAL-STATISTICS-BASED SPATIAL WEIGHT MATRIX INTO A SPATIAL REGRESSION MODEL TO MAP THE DISTRIBUTION OF NON-NATIVE INVASIVE *ROSA MULTIFLORA* IN THE UPPER MIDWEST

Weiming Yu[1], Zhaofei Fan[2*], W Keith Moser[3]

[1] *Department of Forestry, Mississippi State University, Starkville, MS, USA*

[2*] *Cor. Author, School of Forestry and Wildlife Sciences, Auburn University, Auburn, AL, USA*

[3] *USDA Forest Service, Rocky Mountain Research Station, Fort Collins, CO, USA*

Abstract. In this study, we extended the spatial weight matrix defined by Getis and Aldstadt (2004) to a more general case to map the distribution of *Rosa multiflora*, an invasive shrub, across the Upper Midwest counties in a spatial lag model (SLM) context. Both the simulation study and the application to the invasion data of invasive *Rosa multiflora* collected in 2005–2006 proved that the modified spatial weight matrix outperforms its original case and the contiguity- and nearest distance- based spatial weight matrices in diagnostic statistics and resultant invasion maps. The geographical distribution of *Rosa multiflora* in the Upper Midwest was significantly associated with latitude; local clusters (groups of counties) of high abundance/presence of *Rosa multiflora* were significantly determined by TRPF (a ratio of road density to percentage of forest cover at the county level), a variable reflecting the intensity of human disturbance. Both the multiple linear regression model and the SLM models with the original spatial weight matrix and contiguity- and nearest distance- based spatial weight matrices incorporated tended to underestimate the effect of forest type (community) on multiflora rose. As a conclusion, the SLM model incorporating the modified spatial weight matrix has potential applications in mapping spatial data with strong clustering patterns and estimating spatial autocorrelation structure and covariate effects in ecological studies.

Keywords: invasive plants, local spatial statistics, spatial autocorrelation, spatial weight matrix

1 Introduction

The invasion and spread of non-native invasive plants (NNIPs) into natural systems is an ecological phenomenon characterized by spatial autocorrelation across a set of spatial scales and organizational levels (Vermeij 1996). Application of spatial models to map patterns of NNIPs and to quantify contributing factors is of theoretical and applied importance to the monitoring and control of NNIPs. Spatial weight matrices have been used in regression models to account for autocorrelation in data for scientific inferences (e.g., Legendre 1993, Fortin et al. 2006, Kissling and Carl 2008, Zhang et al. 2005, 2009, Lu and Zhang 2010, 2011). Essentially, the purpose of spatial weight matrices is to define the spatial autocorrelation structure of underlying bioecological processes such as the spread of NNIPs across a landscape. Misspecification of spatial autocorrelation usually has two consequences:

1. the estimation of coefficient variances has a downward bias, which results in an inflated type-I error and incorrect inferences; and
2. the estimation of parameters in statistical models is incorrect and results in the wrong interpretation of the environmental variables (Anselin and Bera 1998, Keitt et al. 2002, Haining 2003).

Thus, the choice of a spatial weight matrix is critical for spatial regression models.

It is still a challenge as there are no specific guidelines or schemes for the selection of spatial weight matrices. Stakhovych and Bijmolt (2009) summarized the literature concerning the choice of spatial weight matrices into three avenues of investigation. First, the most popular approach is distance- or neighborhood-related, such as spatial contiguity, inverse distance raised to a certain power, n-nearest neighbors, share of common boundaries, ranked distance, and centroids (Anselin and Bera 1998,

Waller and Gotway 2004). Anselin (1988) argued that spatial weight matrices should be exogenous and should be based on theoretical assumptions on the spatial structure. However, this approach has its limitations: the spatial weight matrix may not reflect the real spatial structure; and specification of the matrix is model-based. Second, specification of spatial weight matrices is model-based. LeSage and Parent (2007) and Holloway and Lapar (2007) used the Bayesian model to choose the spatial weight matrix. Kostov (2010) used the component-wise model boosting algorithm (Buhlmann 2006) when dealing with the selection of spatial weight matrices. The limitation of the model-based approach is the large number of potential spatial weight matrices and relatively limited computational capability, especially when the number of observations is large. Third, specification of spatial weight matrices is data-driven. Researchers construct spatial weight matrices based on the extracted information about the spatial relationships from existing data (Getis and Aldstadt 2004). Getis and Aldstadt (2004) constructed a spatial weight matrix by using the local spatial statistic Getis-Ord $G_i^*(d)$. Subsequently, Aldstadt and Getis (2006) used a sophisticated algorithm to construct a spatial weight matrix that depended on the local spatial statistic Getis-Ord G_i^* and identified the shape of spatial clusters.

Getis and Aldstadt (2004) found that the spatial weight matrix based on the local spatial statistic performs better than spatial weight matrices based on contiguity, inverse distance, or semi-variance model according to the Akaike Information Criterion (AIC) (Akaike 1974). They attributed these results to the local adaptive nature of this spatial weight matrix. However, this weight matrix is not directly related to the distance even though the local spatial statistic Getis-Ord $G_i^*(d)$ is implicitly related to the distance. According to Tobler's first law (Tobler 1970), the inverse distance is used to weight the local spatial statistic $G_i^*(d)$ in this study such that the modified spatial weight matrix is explicitly related to the distance while maintaining its local adaptive nature. Then a natural and immediate question is: Which matrix will perform best?

The major objective of this study is 1) to modify the spatial weight matrix defined by Getis and Aldstadt (2004), and compare the performance of the modified spatial weight matrix and its original case through a simulation study; and 2) compare the performance of the original and modified (local-statistics based) spatial weight matrix with the commonly used contiguity- based and nearest distance-based spatial weight matrix through an application to field data to map the distribution of *Rosa multiflora*, a major non-native invasive shrub, in the Upper Midwest states. In addition, this study will explore the effect of incorporated spatial weight matrices on covariate selection and interpretation through comparison of the derived spatial regression models and a multiple linear regression model, a reference model that does not incorporate a spatial weight matrix.

2 Definition of Getis-Ord $G_i^*(d)$

Getis and Ord (1992) and Ord and Getis (1995) introduced the spatial statistics $G(d)$, $G_i(d)$ and $G_i^*(d)$. $G(d)$ is a global indicator of spatial clustering; but $G_i(d)$ and $G_i^*(d)$ can be used to detect local clusters. These three statistics are defined as,

$$G(d) = \frac{\sum_i \sum_j w_{ij}(d) x_i x_j}{\sum_i \sum_j x_i x_j}, \tag{1}$$

$$G_i(d) = \frac{\sum_j \left(w_{ij}(d)x_j - W_i \bar{x}(i)\right)}{s(i)\{[(nS_{1i}) - W_i^2]/(n-2)\}^{1/2}},\ j \neq i, \text{and} \tag{2}$$

$$G_i^*(d) = \frac{\sum_j \left(w_{ij}(d)x_j - W_i^* \bar{x}\right)}{s * \{[(nS_{1i}^*) - W_i^{*2}]/(n-1)\}^{1/2}},\ all\ j \tag{3}$$

Where, $w_{ij}(d)$ is a symmetric 0 or 1 spatial weight matrix with 1 for all links defined as being within distance d of the i^{th} observation; $W_i = \sum_{j \neq i} w_{ij}(d), W_i^* = W_i + w_{ii}, S_{1i} = \sum_j w_{ij}^2, j \neq i, \text{and } S_{1i}^* = \sum_j w_{ij}^2$; $\bar{x}(i) = \frac{\sum_j x_j}{(n-1)}$, and $s^2(i) = \frac{\sum_j x_j^2}{(n-1)} - (\bar{x}(i))^2$, $j \neq i$; $\bar{x}$ and s^2 denote the usual sample mean and variance, respectively.

A positive value of $G_i^*(d)$ indicates a cluster of relatively high values within d of the i^{th} observation; a negative value of $G_i^*(d)$ indicates a cluster of relatively low values within d of the i^{th} observation. The difference between $G_i(d)$ and $G_i^*(d)$ is that the former does not consider the contribution of the i^{th} observation, but the latter does.

2.1 Constructing spatial weight matrices using Getis-Ord $G_i^*(d)$

In general, $G_i^*(d)$ values monotonically increase around the i^{th} observation as the distance from that observation increases up to a point, then begin to decrease. This point is defined as the critical distance, d_c, (Getis and Aldstadt 2004), where any continuity in spatial dependence or association over distance ends; thus, it defines the cluster diameter (Getis and Aldstadt 2004). To compute $G_i^*(d)$, we need to define the neighbors of the i^{th} observation. Getis and Aldstadt (2004) calculated d_c based on one unit separating centers of rook's case neighbors – the neighbors share a common boundary – within the distance d_c. For simplicity, in this study, we use all neighbors within d_c. We also denote d_1 as the distance to the first nearest neighbor. Then Getis

Table 1: Results of simulation study, where SLM 1 uses the spatial weight matrix $W*$ and SLM 2 uses the modified spatial weight matrix $W**$.

		AIC		Estimated ρ		Moran's I of residuals	
		SLM 1	SLM 2	SLM 1	SLM 2	SLM 1	SLM 2
Random N=25	Mean	2486.6	2486.8	0.78	0.78	0.35	0.35
	Max	2613.6	2613.3	0.88	0.88	0.42	0.42
	Min	2404.3	2404.5	0.67	0.67	0.29	0.29
	SD	56.2	56.2	0.05	0.05	0.03	0.03
2-Cluster N=25	Mean	1677.9	1663	1.17	1.13	1.09	1.14
	Max	1796.1	1789.5	1.2	1.15	1.19	1.24
	Min	1550.5	1533.2	1.13	1.09	1.04	1.09
	SD	64.5	66.8	0.02	0.02	0.03	0.03
6-Cluster N=25	Mean	1420.2	1400.1	1.18	1.14	1.03	1.07
	Max	1492.8	1481.9	1.2	1.16	1.06	1.11
	Min	1318.8	1295.1	1.14	1.11	0.99	1.04
	SD	47.7	49.8	0.02	0.01	0.02	0.02

and Aldstadt (2004) defined the spatial weight matrix W^* as,

1. When $d_c = 0$,

$$w^*_{ij} = 0, \text{ for all } j$$

2. When $d_c = d_1$,

$$w^*_{ij} = 1, \text{ for all } j, \text{ where } d_{ij} = d_c;$$
$$w^*_{ij} = 0, \text{ otherwise;}$$

3. When $d_c > d_1$,

$$lw^*_{ij} = \frac{|G^*_i(d_c) - G^*_i(d_{ij})|}{|G^*_i(d_c) - G^*_i(0)|}, \text{ for all } j \text{ where } d_{ij} \leq d_c;$$
$$w^*_{ij} = 0, \text{ otherwise.} \quad (4)$$

Where, $G^*_i(d_c)$ is the G^*_i score at d_c, and $G^*_i(0)$ is the G^*_i score for the i^{th} observation only and $G^*_i(0)$ is the base from to which other measures of $G^*_i(d)$ are compared. According to the definition, w^*_{ij} is 0 for all the observations that have no spatial correlation with its neighbors, including that d_c is equal to 0 or the distance between the i^{th} and j^{th} observation is greater than d_c.

According to the definition of Getis-Ord $G_i{}^*(d)$, the distance (d) identifies which observations should be included in the calculation, and the statistic $G_i{}^*(d)$ is implicitly related to the distance. Thus, W^* is implicitly related to the distance. However, Carl and Kühn (2007) argued that the similarity of the values of two observations is inversely related to the geographical distance between each other. In other words, we should assign greater weights to the observations closer to the given observations and smaller weights to the observations that are farther away from the given observations. Therefore, in this study we used inverse-distance to weight the Getis–Ord $G_i{}^*(d)$ in order to obtain a modified spatial weight matrix W^{**} such that it explicitly reflects the impact of distance and meanwhile maintains the local adaptive nature of W^*, thereby outperforming the other spatial weight matrices. The modified spatial weight matrix W^{**} is defined as,

1. When $d_c = 0$,

$$w^*_{ij} = 0, \text{ for all } j$$

2. When $d_c = d_1$,

$$w^*_{ij} = 1, \text{ for all } j, \text{ where } d_{ij} = d_c;$$
$$w^*_{ij} = 0, \text{ otherwise;}$$

3. When $d_c > d_1$,

$$lw^*_{ij} = \frac{\left|G^*_i(d_c)/d^k_c - G^*_i(d_{ij})/d^k_{ij}\right|}{\left|G^*_i(d_c)/d^k_c - G^*_i(0)\right|}, \text{ for all } j \text{ where :}$$
$$d_{ij} \leq d_c;$$
$$w^*_{ij} = 0, \text{ otherwise.} \quad (5)$$

Where, k is a non-negative constant. For simplicity, in this study, k is set to one.

The Spatial Lag Model (SLM) was used in the simulation study as follows:

$$Y = \alpha + \rho WY + \beta X + \varepsilon, \quad (6)$$

where Y is the response variable generated by the simulation, ρ represents the dependence structure of the variable Y, β represents the effect on the observations that are not correlated with any of their neighbors, and W is a

spatial weight matrix (W^* or W^{**}). Just as in Getis and Aldstadt (2004), we added a dummy variable X, which takes on the value of one for all observations having no dependence structure and zero otherwise, to compensate for the zero-rows effects in W. The parameters were estimated by using maximum-likelihood methods.

2.2 Simulation design For the simulation study, we used the same design as Getis and Aldstadt (2004) except that we changed the cluster radius in the 2-cluster case to 6 instead of 8 to avoid overlap of cells (Table 1). Each of three types of 30 by 30 raster data sets was simulated for 25 times. The 25 replications with cluster patterns were considered to be sufficient for representing a wide variety of spatial structures (Getis and Aldstadt 2004). The three types are a random normal representing a pattern with no spatial autocorrelation among the values placed in the cells, a pattern of 2 clusters of equal sizes, and a pattern of 6 clusters indicating the spatial structure with random combinations of multiple patches of varying sizes. All the values put in the cells were generated from a standard normal distribution. Figure 1 shows one realization of the random normal pattern, 2-cluster pattern, and 6-cluster pattern representing the potential spatial structure of collected data such as the distribution of invasive shrubs *per se*. For the data sets shown in Figure 1, Figure 2 shows the spatial distribution of the critical distances, which determine the size of spatial weights. For the i^{th} observations, the critical distance is calculated as follows: first, for each given d, we found the nearest neighbors of the i^{th} observations within d; second, we calculated the $G_i^*(d)$; last, d_c is defined as the value d such that $G_i^*(d)$ is the maximum and $G_i^*(d)$ is monotonically increasing in the interval (0, d).

2.3 Evaluation criteria Per Getis and Aldstadt (2004), AIC, autocorrelation coefficient ρ, and Moran's I of residuals were used to evaluate the model performance. Given a set of candidate models for the data, the model with a minimum AIC value will be chosen. AIC penalizes the model with more parameters since the value of AIC increases as the number of parameters in the model increases.

Getis and Aldstadt (2004, page 98) argued that "the autocorrelation coefficient gives an interpretation for the possible association between WY and Y". If $\rho = 1$, it means that W is a good representation of the spatial autocorrelation among data; otherwise, if ρ is close to 0, it means that W is not a good representation of the spatial autocorrelation among data. In addition, Getis and Aldstadt (2004) used Moran's I to detect the spatial autocorrelation among the residuals. The Moran's I is

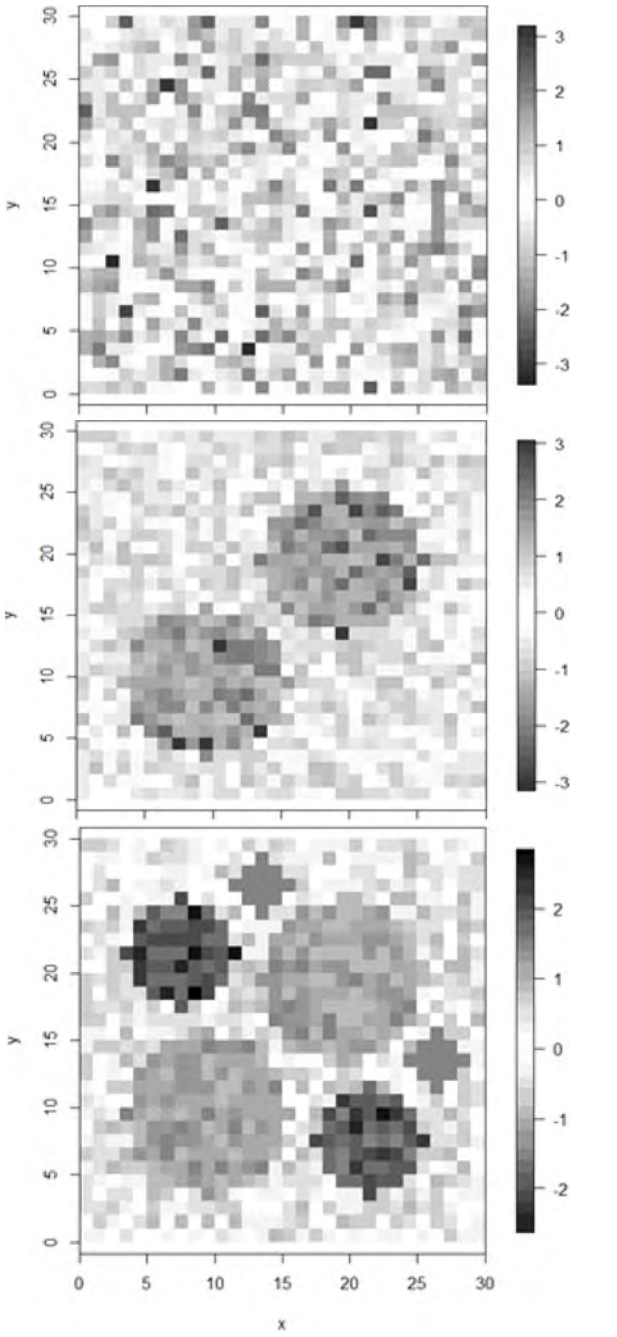

Figure 1: (a) Random data set. 900 values are generated from the standard normal distribution and randomly assigned to the cells of the 30 by 30 grid; (b) 2-cluster data set. 900 values are generated from the standard normal distribution and assigned randomly to 2 clusters: 1 of high value and 1 of low value with radius 6 and centered at (10, 10) and (20, 20); (c) 6-cluster data set. 900 values are generated from the standard normal distribution and assigned randomly to 6 clusters: 3 of high values and 3 of low values with radii 2, 4, and 6 respectively and centered at (14, 27), (27, 14), (8, 22), (22, 8), (10, 10), (20, 20).

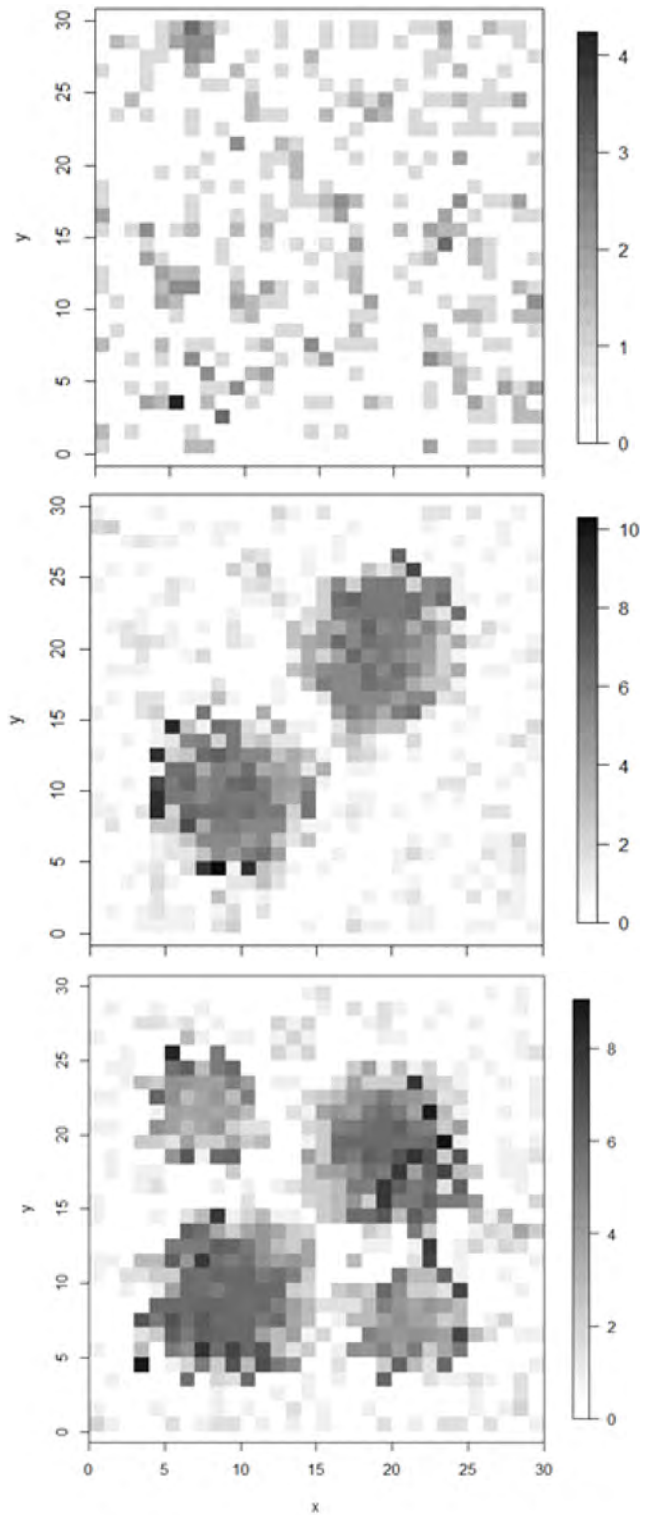

Figure 2: Critical distance (dc) for the random pattern (a), 2-cluster pattern (b), and 6-cluster pattern (c) corresponding to the data sets in Figure 1. Distances are based on nearest neighbors.

defined as,

$$I = \frac{N * \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(e_i - \bar{e})(e_j - \bar{e})}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} * \sum_{i=1}^{N} (e_i - \bar{e})} \quad (7)$$

where, N is the number of cells, e is the residuals vector, and w_{ij} is a spatial weight between the i^{th} observation and j^{th} observation. If W accounts for all of the spatial variation in y, the residuals should be spatially random. Moran's I of the residuals was computed by using the same W that was used in the corresponding candidate models.

An applied example with non-native invasive plant data

Data on non-native invasive plants in forested ecosystems were collected from seven states in the Upper Midwest: Indiana, Illinois, Iowa, Missouri, Michigan, Wisconsin, and Minnesota. Of these states, northern Minnesota, northern Wisconsin, northern Michigan, and southern Missouri are the most heavily forested areas (forested lands >30%), while the middle of this region, covered mostly by prairie prior to European-American settlement, is currently a mosaic of agricultural lands and urban areas (forested lands ˜10%). Extensive human activities (e.g., timber harvesting, clearing land for settlement) in the Upper Midwest created many opportunities for the establishment and spread of NNIPs.

During the 2005–2006 inventory years, 8, 662 U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis (FIA) phase 2 plots in the seven states were assessed for presence or absence and cover (percentage of total plot area) of any of the 25 non-native invasive plant species (Fan et al. 2013). In total, 594 of 649 counties in the Upper Midwest have FIA plots on forest land and 2, 039 (23.5) of 8, 662 FIA plots were invaded by one or more of these invasive shrubs. Among the assessed NNIPs, multiflora rose (*Rosa multiflora Thunb.,* MFR hereafter) was most prevalent and was present in 1, 320 (15.2% of all FIA plots).

Mapping the distributional pattern of MFR at the county level

In this data set, the average number of forested FIA plots per county is 14, but varies dramatically by county and subsequently may affect the estimation of the abundance (presence probability) of MFR. To overcome this potential bias from the sample size, one common way is using the nearest neighbor (moving window) method to adjust the abundance of MFR in a county as:

$$Abundance_i = \frac{\sum_{j \in \eta_i} s_j}{\sum_{j \in \eta_i} n_j} \quad (8)$$

Where, s_j is the number of the presence plots in the county j, n_j is the total plots in the county j, and η_i is the set of counties that share a boundary with the county i, including the county i. However, because this study is not for predictive purposes, but instead for understanding how different spatial weight matrices work in mapping the distributional pattern of MFR, we ignore the sample size effect. The presence probability of MFR in a county was calculated as the ratio of the number of presence plots to the total number of FIA plots and was assigned a value of zero if no FIA plots were installed in a county. An arcsine transformation was conducted on the presence probability of MFR, and the transformed presence probability was used as the response variable in spatial mapping through regression models.

Based on the literature and exploratory data analyses (Fan et al. 2013), the following covariates were selected to model the abundance of MFR (equation (4)): percentage of forest cover as a whole or by major forest-type group at the county level, road (interstate and state highway) density, and latitude/longitude (using the geometric center of each county). The forest cover type map layer was downloaded from the National Atlas (*www.nationalatlas.gov*). Twenty-five forest cover types were obtained from the Advanced Very High Resolution Radiometer (AVHRR) and Landsat Thematic Mapper (TM) imagery. We combined these forest cover types into 7 forest-type groups: *conifer, oak/pine, oak/hickory, oak/gum/cypress, elm/ash/cottonwood, maple/beech/aspen/birch,* and *non-forest*. Their percentage areas (ratio of each forest cover type to area of each county, %) and total forest percent cover were calculated by using GIS for each Upper Midwest county. The road map layer was obtained from the National Atlas (*www.nationalatlas.gov*) and road densities (the length of interstate and state highway per unit area, km/km^2) were calculated for all Upper Midwest counties by using Spatial Analyst in ArcGIS. Furthermore, we defined a new variable, RDPF, as the ratio of the road density and the county forest percentage. We found that RDPF better describes the intensity of human disturbances and forest fragmentation and is more closely correlated with the abundance of invasive shrubs than either road density or county forest percentage alone.

Overall, abundance of MFR changed linearly with the covariates except for latitude. To reflect the "mound" shaped relationship between abundance and latitude/longitude, the quadratic form of latitude/longitude was included in the process of model construction. We used a multiple linear regression model (MLR, neglecting spatial autocorrelation, used to compare with the SLM) and SLM in equation (4) (SLM, the original and modified spatial weight matrix, one contiguity (queen's rule)-based spatial weight matrix and one k-nearest $[k = 3]$ distance-based spatial weight matrix incorporated to account for spatial autocorrelation) to map the change of abundance of multiflora rose with the selected covariates. Here, Y is the arcsine transformed abundance of MFR; X is a set of covariates consisting of: latitude (quadratic form), forest percentage as a whole and by forest-type groups (conifer, oak/pine, oak/hickory, oak/gum/cypress, elm/ash/cottonwood, maple/aspen/beech/birch), road density, and TRPF; β_is are the regression coefficients to be estimated; ν is the independent error vector and assumed to be normally distributed; ρ is the simultaneous autoregressive error coefficient; and W is the spatial weight matrix. SLM 1 and SLM 2 used W^* as defined by Getis and Aldstadt (2004) and the modified W^{**}, respectively; and SLMs 3 and 4 used the contiguity-based spatial weight matrix and the nearest distance-based spatial weight matrix, respectively. AIC was used to select the best model from the candidates. Further, to assess the goodness-of-fit of both SLMs, we reported the Nagelkerke pseudo-R^2, which is defined as:

$$R^2 = \frac{1 - (L(M_{intercept})/L(M_{full})^{2/N}}{1 - (L(M_{intercept}))^{2/N}} \qquad (9)$$

where M_{full} is the model with predictors, $M_{intercept}$ is the model without predictors, $L(\cdot)$ is the likelihood of the corresponding models, and N is the number of observations. We also evaluated the spatial patterns of residuals by using local Moran's I.

All statistical computation, analysis, and simulation were conducted under the R statistical environment (R Development Core Team 2011). The package *spdep* was used to fit the SLM model and the *sp* and *maps* packages were used to draw graphics (Bivand et al. 2008).

2.4 Results and Discussion The simulation results showed that for all cases, the value of the autocorrelation coefficient, ρ, is significantly different from 0 ($p < 0.0001$) (Table 1). In general, the range of Moran's I is in the interval (-1, 1). In the 2- and 6-cluster cases, however, Moran's I values were slightly greater than 1. This result may have occurred because the number N in equation (7) does not represent the actual number of cells as some observations have no spatial autocorrelation with their neighbors and the corresponding rows in the modified spatial weight matrices are zero.

In the random case, the mean AIC values and the variation of AIC were almost the same for both SLMs. In the 2-cluster case, the mean AIC value of SLM 2 is not significantly different from that of SLM 1. In the 6-cluster case, the mean AIC value of LSM 2 is significantly smaller than that of SLM 1 at 10% significance level ($p = 0.0755$). However, the estimated mean values of the autocorrelation coefficient, ρ, and the Moran's I value of residuals have small but significant differences for both

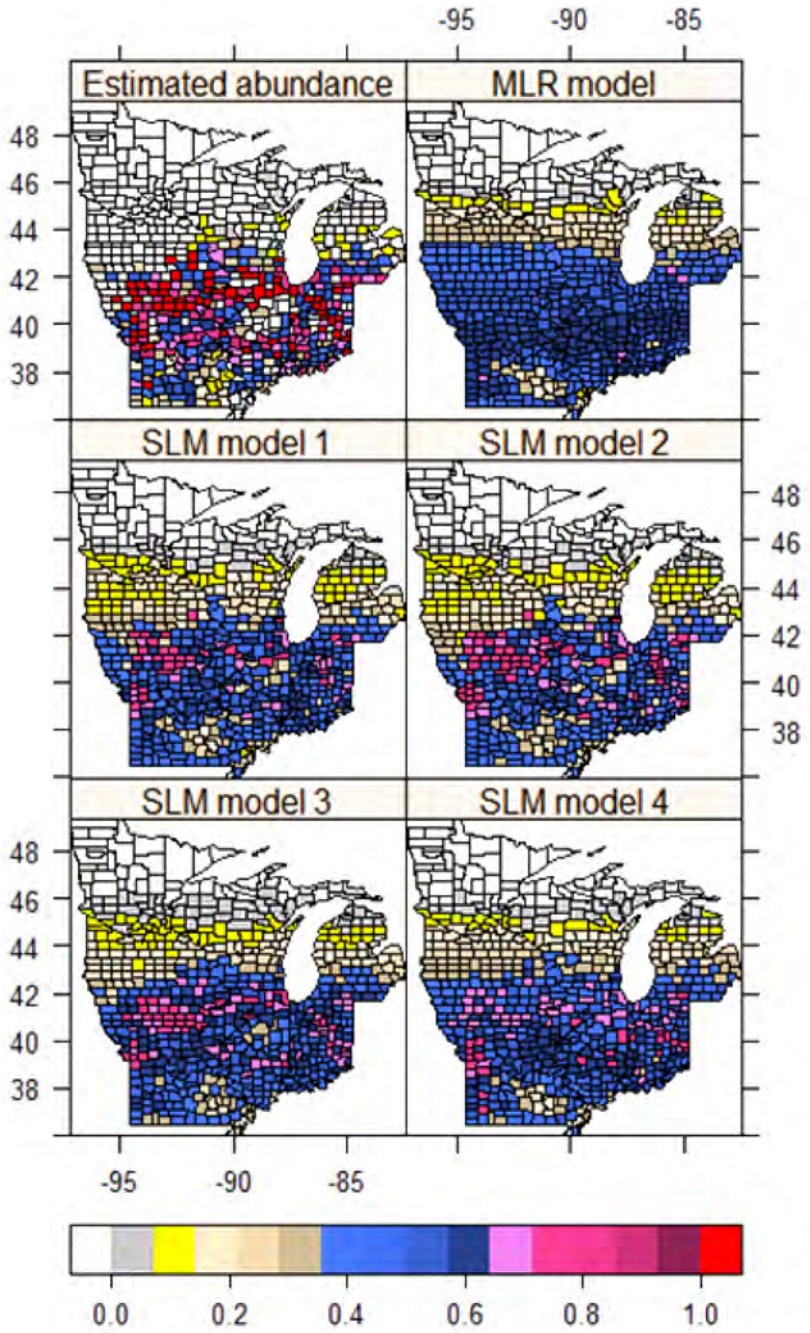

Figure 3: The empirically estimated and model mapped presence probability of multiflora rose in the upper Midwest, 2005-2006. SLM models 1 and 2 use the original and modified Getis-Ord Gi* based spatial weight matrix, and models 3 and 4 use contiguity- based and nearest distance-based spatial weight matrix, respectively.

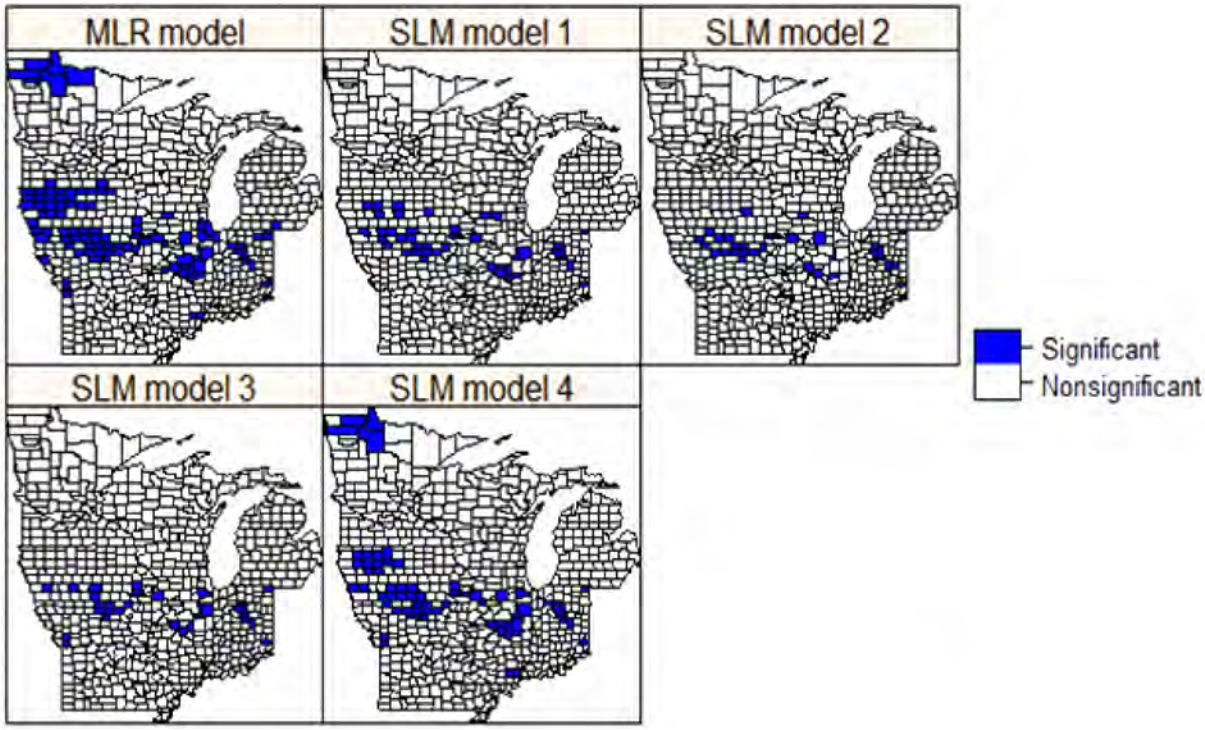

Figure 4: Spatial clusters of residuals of the MLR and SLM models at the significant level of alpha = 0.05. SLM model 2 and model 3 are best for they have fewest and smallest clusters.

models in both the 2- and 6-cluster cases. Thus, we conclude that W^{**} performs better than W^{*} according to the AIC rule when spatial autocorrelation occurred in the data.

Equation (5) can be rewritten as:

$$w_{ij}^{**} = \frac{\left|G_i^*(d_c) - G_i^*(d_{ij}) \times d_c/d_{ij}\right|}{\left|G_i^*(d_c) - G_i^*(0) \times d_c\right|} \qquad (10)$$

In general, if d_{ij} is close to d_c (i.e., d_c/d_{ij} is close to 1), then w_{ij}^{**} is greater than w_{ij}^{*}. If d_{ij} is much smaller than d_c (i.e., d_c/d_{ij} is large enough), then w_{ij}^{**} is smaller than w_{ij}^{*}. Thus, the modification of W^{*} adjusts weights by assigning a greater weight to observations that are far away from i^{th} observations and assigning smaller weights to observations close to i^{th} observations. The better performance of the modified spatial weight matrix may be attributed to the fact that the original case W^{*} over-weighted the observations that are close to i^{th} observations and under-weighted the observations that are far away from i^{th} observations (Carl and Kühn 2007).

Notice that W^{*} is a special case of W^{**} as $k = 0$, and W^{**} extends W^{*} to a more general case without losing its local adaptive property. On the other hand, though k is set to one for simplicity, we may choose an optimal k, which minimizes AIC, by using the trial-and-error method with the data.

The spatial distribution of MFR at the county level showed a strong spatial autocorrelation characterized by several clusters centered in southern Iowa, northeastern Illinois, and north-central Indiana (Figure 3). The SLMs, to varying degrees, captured major clusters; the MLR model, however, failed to distinguish these clusters (Figure 3). Among the SLMs, model 2, which was based on the modified spatial weight matrix (W^{**}), had the smallest AIC and MSE and largest Nagelkerke pseudo-R^2 (Table 2). The spatial patterns of residuals showed that SLM 2 was comparable to SLM 3 (which used the contiguity-based spatial weight matrix), but better than the MLR model and the original Getis-Ord $G_{\text{i}}^{*}(d)$ local statistic and nearest-distance-based SLMs (1 and 4) in terms of fewer and smaller clusters of residuals (Figure 4). A closer county-by-county inspection indicated that the abundance indicated by SLM 2 is larger in most counties in the central and southern parts of the study region, but smaller in the northern portion than for other SLMs. The difference in AIC, MSE, and Nagelkerke pseudo-R^2 values for the MLR model and SLMs can be attributed to the spatial weight matrix since it is the only difference between them. This means the spatial weight matrix based on the modified Getis-Ord $G_{\text{i}}^{*}(d)$ performs better than other spatial weight matrices incorporated in mapping the regional pattern of invasive plants. Fan et

Table 2: The results of multiple linear regression (MLR) model and four spatial lag models (SLMs) to map the presence probability of multiflora rose in the Upper Midwest counties.

Model	Variables	Estimate	SE	p-value	AIC/MSE	*Adjusted R^2
MLR model	Intercept	-24.0244	4.5878	¡0.0000	- /	0.2286
	Latitude	1.2217	0.2182	¡0.0000	0.0893	
	Latitude^2	-0.0151	0.0026	¡0.0000		
	Oak/Hickory	-0.2527	0.1598	0.1142		
	Oak/Gum/Cypress	-7.4027	3.3458	0.0273		
	Elm/Ash/Cottonwood	1.0607	0.5517	0.055		
	RDPF	0.5085	0.1112	¡0.0000		
SLM 1	Intercept	-17.8433	4.7232	0.0002	880.93/	0.2599*
	Latitude	0.9249	0.2245	¡0.0000	0.0723	
	Latitude^2	-0.0116	0.0027	¡0.0000		
	Oak/Hickory	-0.3287	0.1632	0.044		
	Oak/Gum/Cypress	-7.4745	3.2999	0.0235		
	Elm/Ash/Cottonwood	1.2801	0.5918	0.0305		
	RDPF	0.4304	0.1164	0.0002		
SLM 2	Intercept	-15.6433	4.7061	0.0009	863.54/	0.2795*
	Latitude	0.8163	0.2238	0.0003	0.0671	
	Latitude^2	-0.0103	0.0026	¡0.0000		
	Oak/Hickory	-0.314	0.1622	0.0529		
	Oak/Gum/Cypress	-7.1966	3.2533	0.027		
	Elm/Ash/Cottonwood	1.3694	0.5827	0.0188		
	RDPF	0.3983	0.1156	0.0006		
SLM 3	Intercept	-25.5854	5.9048	¡0.0000	869.28/	0.2731*
	Latitude	1.2942	0.2808	¡0.0000	0.0685	
	Latitude^2	-0.016	0.0033	¡0.0000		
	Oak/Hickory	-0.186	0.1904	0.3284		
	Oak/Gum/Cypress	-5.4456	3.5445	0.1245		
	Elm/Ash/Cottonwood	0.8175	0.6213	0.1883		
	RDPF	0.3969	0.1204	0.001		
SLM 4	Intercept	-25.0887	5.0336	¡0.0000	883.01/	0.2575*
	Latitude	1.2732	0.2394	¡0.0000	0.0793	
	Latitude^2	-0.0158	0.0028	¡0.0000		
	Oak/Hickory	-0.2648	0.1697	0.1186		
	Oak/Gum/Cypress	-6.3243	3.5397	0.074		
	Elm/Ash/Cottonwood	1.0099	0.5712	0.077		
	RDPF	0.4209	0.1151	0.0003		

*Nagelkerke pseudo-R^2; SLM 1 and 2 use the original and modified Getis-Ord G_i^* based spatial weight matrix, respectively; and models 3 and 4 use a contiguity-based and nearest distance-based spatial weight matrix, respectively.

al. (2013) mapped the abundance of invasive shrubs and analyzed the associated factors by using the nonparametric kernel smoothing and CART (classification and regression tree), respectively. Invasive shrubs such as MFR were predominantly distributed in the central part of the study area with two major clusters surrounding Chicago, Illinois and Des Moines, Iowa, similar to the simulated 2-cluster pattern. Test statistics indicate that SLM 2 performed well; simulated data conformed well to actual data.

With the model per se, the estimated regression coefficient for RDPF in the SLMs were significantly smaller than that in the MLR model, suggesting that RDPF had less leverage in predicting the abundance of MFR after the spatial autocorrelation was partially interpreted by the spatial weight matrix (Table 2). There was no significant difference in the estimated RDPF among the four SLMs although the estimated values for models 2 and 3 tended to be small compared to models 1 and 4.The estimated regression coefficients for other covariates in models 3 and 4 were nearly identical to those in the MLR model. However, models 1 and 2 had significantly larger estimated values for the intercept and smaller estimated values for latitude (linear term) than the MLR model and SLMs 3 and 4. Further, forest type covariates were statistically significant in models 1 and 2 in contrast with the marginal significance or non-significance in the MRL model and models 3 and 4. The MLR model, as a benchmark, did not account for spatial autocorrelation, which resulted in inaccurate parameter estimation and invasive shrub simulation because of the downward bias in the estimation of coefficient variances (Anselin and Bera 1998, Keitt et al. 2002, Haining 2003). The SLMs with the contiguity-based spatial weight matrix (model 3) and the nearest distance-based spatial weight matrix (model 4) may somewhat improve the estimates for certain covariates (e.g., RDPF). But the SLMs with the local statistics-based spatial weight matrix perform better in mapping the regional patterns of clustered MFR data and estimating covariate effects (Table 2).

Compared to the original Getis-Ord $G_i^*(d)$-based spatial weight matrix (in SLM 1), the modified spatial weight matrix (in SLM 2) renders a relatively large effect for the intercept, forest types, and the quadratic term of latitude, but a small effect for RDPF and latitude. This result means that the modified spatial weight matrix explained more spatial variation (clusters) related to latitude and forest types than the original spatial weight matrix as shown by the spatial pattern of residuals (Figure 4). The significant association ($p < 0.0003$) between abundance of MFR and latitude (including its quadratic form) reflected the unimodal distribution pattern in the latitudinal direction (from south to north) with major clusters of high presence probability centered in southern Iowa, northern Missouri, and northern Illinois (Figure 3). The abrupt decrease of invasive shrubs at latitudes greater than 44°N may suggest that the dispersal of MFR, which is the most prominent of the major invasive shrubs, is limited by the extremely low temperatures in the northern part of the region (Doll 2006, Denight et al. 2008).

The abundance of MFR was negatively associated with the forest-type groups oak/hickory and oak/gum/cypress, and positively related to the forest-type group elm/ash/cottonwood in all models. This relationship indicated that MFR were more likely to invade counties with higher proportions of lowland forests such as elm/ash/cottonwood. MFR was less abundant in bottomland forests (e.g., oak/gum/cypress) or upland forests (e.g., oak/hickory), a result consistent with the traits of the major invasive shrubs (e.g., non-native bush honeysuckles). MFR is most productive in sunny areas with well-drained moist uplands and lowlands; it endures shade, sun, and damp or dry conditions, but does not grow well in standing water or in extremely dry areas (Bowman's Hill Wildflower Preserve 1997, Munger 2002, Doll 2006). The appearance of these three forest-type groups in the MLR and SLM models in predicting the abundance of MFR most likely reflected their prominence (e.g., oak/hickory, elm/ash/cottonwood) in the area (i.e., Iowa, Missouri, Illinois, and Indiana) where IPs were widely distributed (Moser et al. 2016).

In this study area, most of the forests are located in the northern part of Minnesota, Wisconsin, and Michigan and in southern Missouri; the central part of the study region is less forested. Human disturbances and urban development result in decreased forest cover accompanied by increased road density. Road development usually leads to increased forest fragmentation and increased habitat for IPs. By altering overall landscape structure and increasing the ratio of forest edge to total forest area, roads open up a higher proportion of forest area to invasion by non-native plants (Saunders et al. 2002, Mortensen et al. 2009). As road density increases, a higher proportion of the landscape becomes roadside habitat, which tends to be highly invaded by IPs (Gelbard and Belnap 2003). Watkins et al. (2003) documented that prevalence of IPs is negatively associated with the distance to road. Von der Lippe and Kowarik (2007) found that long-distance dispersal of seeds of IPs by vehicles is a routine rather than an occasional mechanism, and dispersal of plants by vehicles will accelerate plant invasions. Moser et al. (2008) found that the presence of MFR is significantly and negatively related with distance to road at the plot level. Compared to highway density and county forest percentage, which measure human disturbances from different perspectives, RDPF provides a synthetic measure of the intensity of human

disturbances. The significant ($p < 0.001$), positive association between the abundance of MFR and RDPF confirms human disturbance as one of the major driving factors for its invasion and spread in the Upper Midwest. Greater human presence reduces native forest ecosystem defenses, by upsetting ecological stability via disturbance and increasing the opportunity for niche capture by invasive plants, and often increases propagule pressure, both passively and actively, thus further increasing abundance of the invaders (Macdonald 1994, Pimentel et al. 2005, Richardson and Pysek 2006).

3 Conclusions

In this study, we compared the performance of two spatial weight matrices, W^* and W^{**}, based on the local spatial statistics Getis-Ord G_i^*. We found that with strong spatial autocorrelation of invasive plants such as multiflora rose W^{**} performs better than W^* and the contiguity- and nearest distance-based spatial weight matrix. The spatial weight matrix W^* is a special case of the spatial weight matrix W^{**}. The modified spatial weight matrix W^{**} extends the spatial weight matrix W^* to a more general case without losing its local adaptive property. The SLM with a spatial weight matrix based on local spatial statistics such as Getis-Ord G_i^* provides a promising approach to map regional patterns of sophisticated ecological phenomena such as the invasion and spread of invasive plants. Compared to spatial weight matrices defined through global statistics such as geostatistical or inverse-distance models, the local statistics-based spatial weight matrix demonstrates its strength and efficacy in explaining the spatial variation of the data, particularly when the distribution of invasive plants is controlled by a number of multi-scale processes and takes a spatial pattern of multiple clusters of various sizes.

Our results also indicated that spatial weight matrices might affect the choice of the factors used to explain the abundance of invasive plants. In addition to spatial autocorrelation, which was explained by spatial weight matrices, the variation in abundance of invasive plants such as multiflora rose in the Upper Midwest counties was attributable to human disturbances (RDPF), a geographic factor (latitude), and forest conditions. But the significance of these factors in the model may change with the choice of spatial weight matrices. In this study, with the original and modified spatial weight matrix based on the local spatial statistics Getis-Ord G_i^*, county-level proportions of oak/hickory, oak/gum/cypress, and elm/ash/cottonwood forest-type groups became significant at the significance level of $\alpha = 0.05$ compared to the marginal significance or non-significance with the MLR model and the SLMs with contiguity- and distance-based spatial weight matrix. The distribution of multiflora rose was primarily related to latitude and human disturbance, but forest types were also related to the observed clustering patterns. The SLM with the modified spatial weight matrix performs better in explaining the spatial patterns of multiflora rose and detecting secondary covariate effects than the MLR model and the SLMs with spatial weight matrices based on the original Getis-Ord G_i^* contiguity, or distance.

References

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 19(4): 716–723.

Aldstadt, J., and A. Getis. 2006. Using AMOEBA to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*. 38: 327–343.

Anselin, L. 1988. *Spatial Econometrics: Method and Models*. the Netherlands: Kluwer Academic Publishers.

Anselin, L., and A. Bera. 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In *Handbook of Applied Economic Statistics*, Ullah. A. (ed.). Marcel Dekker, New York.

Bivand, R. S., E. J. Pebesma, and V. Gomez-Rubio. 2008. *Applied Spatial Data Analysis with R*. Springer, New York.

Bowman's Hill Wildflower Preserve (BHWP). 1997. *Multiflora rose - invasive exotics.* Available online at: www.bhwp.org/cms/files/file_ID96816.pdf; last accessed: Sept. 29, 2011.

Buhlmann, P. 2006. Boosting for high-dimensional linear models. *The Annals of Statistics*. 34(2): 559–583.

Carl, G., and I. Kühn. 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and Logit models. *Ecological Modelling*. 207(2-4): 159–170.

Denight, M. L., P. J. Guertin, D. L. Gebhart, and L. Nelson. 2008. *Invasive Species Biology, Control, and Research, Part 2: Multiflora Rose (Rosa Multiflora)*. Available online at www.cecer.army.mil/techreports/ERDC_TR-08-11/ERDC_TR-08-11.pdf; last accessed Sept. 29, 2011.

Doll, J. D. 2006. Biology of Multiflora Rose. P. 239 in *North Central Weed Science Society Proceedings*. Available online at www.ncwss.org/proceed/2006/abstracts/239.pdf. last accessed Sept. 29, 2011.

Fan, Z., W. K. Moser, M. H. Hansen, and M. D. Nelson. 2013. Regional patterns of major non-native invasive plants and associated factors in Upper Midwest forests. *Forest Science.* 59 (1): 38–49.

Fortin, M.J., M.R.T. Dale, and J. Hoef. 2006. Spatial analysis in ecology. *Encyclopedia of Environmetrics.* 4: 2051–2058.

Gelbard, J. L. and J. Belnap. 2003. Roads as conduits for exotic plant invasions in a semiarid landscape. *Conservation Biology.* 17(2): 420–432.

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis.* 24(3): 189–206.

Getis, A., and J. Aldstadt. 2004. Constructing the spatial weights matrix using a local statistic. *Geographical Analysis.* 36(2): 90–104.

Haining, R. P. 2003. *Spatial Data Analysis: Theory and Practice.* Cambridge University Press, Cambridge, United Kingdom.

Holloway, G., and M. L. A. Lapar. 2007. How big is your neighborhood? Spatial implications of market participation among Filipino smallholders. *Journal of Agricultural Economics.* 58(1): 37–60.

Keitt, T. H., O. N. Bjørnstad, P. M. Dixon, and S. Citron-Pousty. 2002. Accounting for spatial pattern when modeling organism-environment interactions. *Ecography.* 25(5): 616–625.

Kissling, W. D., and G. Carl. 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography.* 17(1): 59–71.

Kostov, P. 2010. Model boosting for spatial weighting matrix selection in spatial lag models. *Environment and Planning B: Planning and Design.* 37(3): 533–549.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology.* 74(4): 1659–1673.

LeSage, J. P., and O. Parent. 2007. Bayesian model averaging for spatial econometric models. *Geographical Analysis.* 39(3): 241–267.

Lu, J., and L. Zhang. 2010. Evaluation of parameter estimation methods for fitting spatial regression models. *Forest Science* 56(5): 505–514.

Lu, J., and L. Zhang. 2011. Modeling and prediction of tree height-diameter relationships using spatial autoregressive models. *Forest Science.* 57 (3): 252–264.

Macdonald, I. A. W. 1994. Global change and alien invasions: Implications for biodiversity and protected area management. P. 197-207 In: *Biodiversity and Global Change*, Solbrig, O. T., P. G. van Emden, and W. J. van Oordt. Wallingford-Oxon, UK: CAB International.

Mortensen, D. A., E. Rauschert, A. N. Nord, and B. P. Jones. 2009. Forest roads facilitate spread of invasive plants. *Invasive Plant Science and Management.* 2(3): 191–199.

Moser, W. K., M. H. Hansen, M. D. Nelson, and M. H. William. 2008. Relationship of invasive groundcover plant presence to evidence of disturbance in the forests of the upper Midwest of the United States. P. 29-58 in *Invasive Plants and Forest Ecosystems*, Kohli, R.H., S. Jose, H. P. Singh, and D. R. Batish. CRC Press, Boca Raton, FL.

Moser, W.K., Z. Fan, M.H., Hansen, M.K. Crosby, S.X. Fan. 2016. Invasibility of three major non-native shrubs and associated factors in Upper Midwest U.S. forest lands. *Forest Ecology and Management.* 379: 195-205.

Munger, G. T. 2002. Rosa multiflora. In: *Fire Effects Information System.* U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory; Available online at www.fs.fed.us/database/feis/; last accessed Sept. 29, 2011.

Ord, J. K., and A. Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis.* 27(4): 286–306.

Pimentel, D., R. Zuniga, and D. Morrison. 2005. Update on the environmental and economic costs Associated with alien-invasive species in the United States. *Ecological Economics.* 52(3): 273–288.

R Development Core Team. 2011. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/.

Richardson, D. M., and P. Pysek. 2006. Plant invasions: merging the concepts of species invasiveness and community invisibility. *Progress in Physical Geography.* 30: 409–431.

Saunders, S. C., M. R. Mislivets, J. Chen, and D. T. Cleland. 2002. Effects of roads on landscape structure within nested ecological units of the northern Great Lakes region, USA. *Biological Conservation.* 103(2): 209–225.

Stakhovych, S., and T. H. A. Bijmolt. 2009. Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science.* 88(2):389–408.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography.* 46:234–240.

Vermeij, G.J. 1996. An agenda for invasion biology. *Biological Conservation.* 78(1-2): 3–9.

Von der Lippe, M., and I. Kowarik. 2007. Long-distance dispersal of plants by vehicles as a driver of plant invasions. *Conservation Biology.* 21(4): 986–996.

Waller, L. A., and C. A. Gotway. 2004. *Applied Spatial Statistics for Public Health Data.* John Wiley & Sons, Inc.

Watkins, R. Z., J. Chen, J. Pickens, and K. D. Brosofske. 2003. Effects of roads on understory plants in a managed hardwood landscape. *Conservation Biology.* 17(2): 411–419.

Zhang, L., J. H. Gove, and L. S. Heath. 2005. Spatial residual analysis of six modeling techniques. *Ecological Modelling.* 186: 154–177.

Zhang, L., Z. Ma, and L. Guo. 2009. Spatial autocorrelation and heterogeneity in the relationships between tree variables. *Forest Science.* 55(4): 533–548.