Constructing Spatial Weight Matrix Using Local Spatial Statistics And Its Applications

Weiming Yu

Follow this and additional works at: https://scholarsjunction.msstate.edu/td

Recommended Citation
https://scholarsjunction.msstate.edu/td/1164
CONSTRUCTING SPATIAL WEIGHT MATRIX USING LOCAL SPATIAL STATISTICS AND ITS APPLICATIONS

By

Weiming Yu

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Forestry
in the Department of Forestry

Mississippi State, Mississippi

December 2011
Copyright 2011

By

Weiming Yu
CONSTRUCTING SPATIAL WEIGHT MATRIX USING LOCAL SPATIAL
STATISTICS AND ITS APPLICATIONS

By

Weiming Yu

Approved:

Zhaofei Fan
Assistant Professor of Forestry
(Major Professor)

William H. Cooke III
Associate Professor of Geosciences
(Committee Member)

Changyou Sun
Associate Professor of Forestry
(Committee Member)

Andrew W. Ezell
Professor of Forestry
Head and Graduate Coordinator
of the Department of Forestry

George M. Hopper
Dean
College of Forest Resources
In this study, we extend the spatial weight matrix defined by Getis and Aldstadt (2004) to a more general case. The modified spatial weight matrix performs better than the original spatial weight matrix since the modified spatial weight matrix adjusts weights of observations based on the distance from other observations. Both the simulation study and the application to the ecological process of invasion of non-native invasive plants (NNIPs) provide evidences for the better performance of the modified spatial weight matrix. We also develop procedures that can be used to quantify the invasion stages of NNIPs. The resultant map of invasion stage on county-level provides a useful and meaningful tool for policy makers; especially, it can be used to optimize allocation of management resources. The result of simultaneous autoregressive model shows that not only the biotic and abiotic factors but also human activities play an important role in the establishment and spread of multiflora rose in the Upper Midwest. It also shows the tendency of the establishment and spread of multiflora rose (Rosa Multiflora, Thunb. ex Murr.) in the Upper Midwest.
Key words: local spatial statistics, Moran’s I, multiflora rose, non-native invasive plants, spatial autocorrelation, spatial autoregressive model, spatial dependence, spatial weight matrix
DEDICATION

I would like to dedicate this research to my wife, Qian Wang and my parents.
ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my major advisor, Dr. Zhaofei Fan, for his patience, guidance and assistance throughout my graduate program. I am also grateful for the valuable assistance of my committee members, namely, Dr. William H. Cooke III and Dr. Changyou Sun. I extend my appreciation to Dr. W. Keith Moser from the US Forest Service Northern Research Station for conceptualizing the original project and for providing funding and data for this study and to the Department of Forestry at Mississippi State University for providing an excellent study environment. I thank my classmates and friends for their help and encouragement. Finally, I thank my wife and my family for their endless support and love.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>DEDICATION</strong></td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td><strong>ACKNOWLEDGEMENTS</strong></td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td><strong>LIST OF TABLES</strong></td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td><strong>LIST OF FIGURES</strong></td>
<td>vii</td>
</tr>
<tr>
<td>I</td>
<td><strong>INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Problem Statement</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Objectives</td>
<td>2</td>
</tr>
<tr>
<td>II</td>
<td><strong>AN EXTENTION OF CONSTRUCTING THE SPATIAL WEIGHT MATRICES USING LOCAL SPATIAL STATISTICS</strong></td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Spatial Autocorrelation</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>The choice of Spatial Weight matrices</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Objective</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Method</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Definition of Getis-Ord $G_i*(d)$</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Constructing Spatial Weight Matrices Using Getis-Ord $G_i*(d)$</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Simulation Design</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Evaluation Criteria</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Results and Discussion</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>16</td>
</tr>
<tr>
<td>III</td>
<td><strong>DISPERLAL OF MULTIFLORA ROSE IN THE UPPER MIDWEST</strong></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Study Area</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>Methodology</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>The presence probability of MFR</td>
<td>23</td>
</tr>
</tbody>
</table>

iv
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>10</td>
<td>Spatial patterns generated and adapted from Getis and Aldstadt (2004)</td>
</tr>
<tr>
<td>2.2</td>
<td>15</td>
<td>Results of simulation, where LSM 1 uses the spatial weight matrix $W^*$ and LSM 2 uses the modified spatial weight matrix $W^{**}$</td>
</tr>
<tr>
<td>3.1</td>
<td>22</td>
<td>Forested land in seven Upper Midwest States</td>
</tr>
<tr>
<td>3.2</td>
<td>22</td>
<td>Area of sampled forested land by States and Forest type group (in km$^2$)</td>
</tr>
<tr>
<td>3.3</td>
<td>22</td>
<td>Cover codes and ranges of percent cover of non-native invasive plants used in recording invasive species’ presence, FIA plots, 2005-2006</td>
</tr>
<tr>
<td>3.4</td>
<td>43</td>
<td>Results of SAR models (part of non-significant variables is removed from the models. This table only includes variables that are significant at 10% significance level at least in one model.)</td>
</tr>
<tr>
<td>3.5</td>
<td>45</td>
<td>The final estimation of SAR models (This table only includes variables that are significant at 10% significance level at least in one model.)</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Random Data Set. 900 values are generated from the standard normal distribution and randomly assigned to the cells of the 30 by 30 grid.</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>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).</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>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 (14, 27), (27, 14), (8, 22), (22, 8), (10, 10), (20, 20).</td>
<td>11</td>
</tr>
<tr>
<td>2.4</td>
<td>Critical distance ($d_c$) for the random pattern in Figures 2.1. Distances are based on nearest neighbors.</td>
<td>12</td>
</tr>
<tr>
<td>2.5</td>
<td>Critical distance ($d_c$) for the 2-Cluster pattern in Figures 2.2. Distances are based on nearest neighbors.</td>
<td>12</td>
</tr>
<tr>
<td>2.6</td>
<td>Critical distance ($d_c$) for the 6-Cluster pattern in Figures 2.3. Distances are based on nearest neighbors.</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Part of the distribution of the total number of FIA plots in each county (with less than 7 FIA plots) in the Upper Midwest.</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>The classification procedure of invasion stages based on the estimated presence probability.</td>
<td>30</td>
</tr>
<tr>
<td>3.3</td>
<td>The classification procedure of invasion stages based on the estimated presence probability and cover rate.</td>
<td>31</td>
</tr>
<tr>
<td>3.4</td>
<td>The distribution of FIA Phase 2 Plots without (green) and with the presence of MFR (red) in the Upper Midwest, 2005-2006.</td>
<td>35</td>
</tr>
<tr>
<td>3.5</td>
<td>The smoothed presence probability of MFR in the Upper Midwest, 2005-2006.</td>
<td>36</td>
</tr>
</tbody>
</table>
3.6 The smoothed cover rate of MFR in the Upper Midwest, 2005-2006. ..........36
3.7 The plot of X relative error (relative cross-validation error) vs. the classification tree size (the number of terminal nodes). .........................37
3.8 Classification and regression tree partition of the estimated presence probability of MFR (pp-presence probability). ..........................37
3.9 Proportion of plots vs. the midpoint value of cover class. .........................38
3.10 The map of invasive stages based on the estimated presence probability of MFR in the Upper Midwest, 2005-2006. .................................38
3.11 The plot of X relative error (relative cross-validation error) vs. the classification tree size (the number of terminal nodes). .........................40
3.12 Classification and regression tree partition of the estimated presence probability and estimated cover rate of MFR (pp-presence probability, cr-cover rate). ..........................................................41
3.13 The plot of estimated cover rate vs. estimated presence probability and the result of classification and regression tree partition based on the estimated presence probability and estimated cover rate of MFR (pp-presence probability). .........................................................41
3.14 Proportion of plots vs. the midpoint value of cover class. ..........................42
3.15 The map of invasive stages based on the estimated presence probability and estimated cover rate of MFR in the Upper Midwest, 2005-2006. .........................................................................................42
CHAPTER I
INTRODUCTION

Problem Statement

In the last few decades, one of the most important and sometimes unavoidable ecological concerns is to identify and explain spatial autocorrelation (some studies used the term spatial dependence; see Fortin and Dale 2005) presented in ecological phenomena. Spatial regression models used in ecological studies take account of spatial autocorrelation by employing a spatial weight matrix. Essentially, the choice of spatial weight matrices is to define the spatial autocorrelation structure of underlying ecological processes. Thus, the choice of a spatial weight matrix is critical for spatial regression models. However, it is still a challenge since there are no specific guidelines and/or schemes about the selection of spatial weight matrices.

Getis and Aldstadt (2004) constructed a spatial weight matrix using the local spatial statistics Getis-Ord $G_i^*(d)$. They found that this spatial weight matrix performs best compared with the spatial weight matrices based on contiguity, inverse distance, or semi-variance model according to the Akaike Information Criterion (AIC). They attributed it to the fact that this spatial weight matrix is locally adaptive. However, this spatial 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 the Tobler’s first law (1970), the inverse distance is used to weight the local spatial statistics $G_i^*(d)$ such that the modified spatial weight matrix is explicitly related to the distance. Then a
natural and immediate question is which one will perform better? In this study, we compare the performance of these spatial weight matrices based on the local spatial statistic through both a simulation study and an application to the dispersal of Multiflora Rose in seven Upper Midwest states.

On the other hand, Theoharides and Dukes (2007) qualitatively classified invasive process of non-native plants into four stages: introduction, colonization, establishment and spread. Webster et al. (2006) found that efforts to control invasive plants should focus on the establishment or earlier stages. But how to quantify the invasion stages is still a practical and meaningful problem since it is related to optimistically reallocate limited resources in forest management.

**Objectives**

The primary objective of this thesis research project is to modify the spatial weight matrix defined by Getis and Aldstadt (2004) to obtain a new spatial weight matrix with better performance. Chapter II compares the performance of the two spatial weight matrices through a simulation study. Chapter III compares the performance of the two spatial weight matrices in Spatial Autoregressive Model (SAR) models to identify the driving factors of the multiflora rose (MFR) dispersal in the seven Upper Midwest states, develop procedures that can be used to classify the invasion stage of invasive plants. Chapter IV summarizes the major results and findings of this study.
CHAPTER II
AN EXTENTION OF CONSTRUCTING THE SPATIAL WEIGHT MATRIX USING LOCAL SPATIAL STATISTICS

Introduction

Spatial Autocorrelation

Tobler’s (1970) first law stated that “everything is related to everything else, but near things are more related than distant things”. In other words, spatial autocorrelation “means that observations close to each other geographically are more likely to be similar than those far away from each other” (Carl and Kühn 2007). Fortin and Dale (2005) classified spatial autocorrelation into two types: inherent spatial autocorrelation and induced spatial autocorrelation. Inherent spatial autocorrelation results from endogenous processes such as species growth, reproduction, migration, and mortality; induced spatial autocorrelation results from exogenous processes such as geomorphologic processes on land, wind, and climate change (Legendre 1993).

Spatial autocorrelation has three important properties. First, spatial autocorrelation may change at different spatial scales since the dominating processes may be different when spatial scale changed from landscape to region, and globe (Legendre 1993). Secondly, the spatial structure of the spatial patterns can be classified into four categories: trend (or gradient), patch (or aggregation, clumping), random and uniform (regular or overdispersed) (Fortin and Dale 2005). The trend and patch are the most common observed in the ecological data since “living beings in nature are distributed...
neither uniformly nor at random” (Legendre 1993). Last, the intensity of spatial autocorrelation may show directionality. The spatial pattern is called as “anisotropic” if the intensity and range of spatial autocorrelation vary with direction or as “isotropic” if spatial autocorrelation intensity varies similarly with distance in all directions (Fortin and Dale 2005). Moran’s I (Moran 1950), Geary’s C (Geary 1954) and Getis-Ord G (Getis and Ord 1992) are widely used to measure the global spatial autocorrelation. Local Moran’s I (Anselin 1995), local Geary’s C and Getis-Ord $G^*_i(d)$ (Getis and Ord 1992; Ord and Getis 1995) are local indicators of spatial autocorrelation.

Legendre (1993) classified two approaches that deal with spatial autocorrelation in the regression models. First, the so-called “raw data approach” introduce a polynomial of geographic coordinates into the model. This approach only allows the large-scale spatial autocorrelation. The second approach is the so-called “matrix approach,” such as simultaneous autoregressive (SAR) models (Waller and Gotway 2004), which take account of spatial autocorrelation by introducing a distance matrix into the model. The matrix approach allows both the small-scale and large-scale spatial autocorrelation to be modeled.

The classical linear regression models assume independently and identically distributed errors (Legendre 1993; Kissling and Carl 2008). However, spatial autoregressive models violate this assumption. Misspecification of spatial autocorrelation usually results in two consequences: 1) the estimation of coefficients variances bias downward, which results in an inflated type-I error and wrong inferences; 2) the estimation of parameters in statistical models is incorrect and results in the wrong interpretation of the environment variables (Anselin and Bera 1998; Keitt et al. 2002; Haining 2003).
The choice of Spatial Weight matrices

The choice of a spatial weight matrix is critical for spatial autoregressive models. However, there are no guidelines about the choice of spatial weight matrices. Stakhovych and Bijmolt (2009) summarized the literatures about the choice of spatial weight matrices into three streams. First, the most popular approach is distance related or neighborhood related, such as spatial contiguity, inverse distance raised to some power, N nearest neighbors, share of common boundaries, ranked distance, centroids, etc. (Anselin and Bera 1998; Waller and Gotway 2004). Anselin (1988) argued that spatial weight matrices should be exogenous and be based on theoretical assumptions on the spatial structure. However, the limitation is that the spatial weight matrix may not reflect the real spatial structure. Secondly, 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 computation capability, especially if the number of observations is large. Last, specification of spatial weight matrices is data-driven. Researchers “allow study data to speak for themselves, that is, they extract from the already existing data whatever spatial relationships appear to be the case and then create a spatial weight matrix from the observed spatial associations” (Getis and Aldstadt 2004). Getis and Aldstadt (2004) constructed a spatial weight matrix using local spatial statistic Getis-Ord $G_i^*(d)$. Aldstadt and Getis (2006) used a sophisticated algorithm to construct a spatial weight matrix which depended on local spatial statistic Getis-Ord $G_i^*$ and identified the shape of spatial clusters.
Objective

The objective of this study is to extend the spatial weight matrix defined by Getis and Aldstadt (2004) to a more general case such that the modified spatial weight matrix is a better representation of spatial autocorrelation. We compare the performance of two spatial weight matrices through a simulation study.

Method

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 respectively defined as:

$$G(d) = \frac{\sum_{j} w_{ij}(d)x_jx_j}{\sum x_jx_j}$$

(2-1)

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

(2-2)

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

(2-3)

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 a given $i$; $W_i^* = W_i + w_{it}$, $S_i = \sum_j w_{ij}^2, j \neq i$, and $S^*_i = \sum_j w^2_{ij}$; $\bar{x}(i) = \frac{\sum_j x_j}{(n-1)}$, and $s^2(i) = \frac{\sum_j x_j^2}{(n-1)} - (\bar{x}(i))^2, j \neq i$; $\bar{x}$ and $s^2$ respectively denote the usual sample mean and variance.
A positive value of $G_i^*(d)$ indicates a cluster of relatively high values within distance $d$ of $i$th observation; a negative value of $G_i^*(d)$ indicates a cluster of relatively low values within distance $d$ of $i$th observation. The difference between $G_i(d)$ and $G_i^*(d)$ is that only $G_i^*(d)$ considers the contribution of $i$th observation when $G_i(d)$ and $G_i^*(d)$ are computed.

**Constructing Spatial Weight Matrices Using Getis-Ord $G_i^*(d)$**

In general, $G_i^*(d)$ values monotonically increase around each observation as the distance increase from it and then $G_i^*(d)$ values decrease after a certain distance, which is defined as the critical distance $d_c$ (Getis and Aldstadt 2004). The critical distance $d_c$, where any continuity in spatial dependence over distance ends, is the cluster diameter (Getis and Aldstadt 2004). To compute $G_i^*(d)$, we need to define the neighbors of $i$th observation, which determine the observations that are used to compute the $G_i^*(d)$. Getis and Aldstadt (2004) calculated the critical distance $d_c$ based on one unit separating centers of rook’s case neighbors – the neighbors share a common boundary – within distance $d$. For simplicity, in this study, we use all neighbors within distance $d$. We also denote $d_1$ as the distance to the first nearest neighbor. Then Getis and Aldstadt (2004) defined the spatial weight matrix $W^*$ as:

1. if $d_c = 0$, $w^*_{ij} = 0$ for all $j$;
2. if $d_c = d_1$, $w^*_{ij} = 1$ for all $j$ where $d_{ij} = d_c$; $w^*_{ij} = 0$, otherwise;
3. if $d_c > d_1$, $w^*_{ij} = \frac{G_i^*(d_c) - G_i^*(d_{ij})}{G_i^*(d_c) - G_i^*(0)}$ for all $j$ where $d_{ij} \leq d_c$; $w^*_{ij} = 0$, otherwise.

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

According to the definition of Getis-Ord $G_i^*(d)$, the distance $d$ play the role that identify which observation should be included in calculation, and the statistics $G_i^*(d)$ is implicitly related with the distance. Thus, the spatial weight matrix $W^*$ is implicitly related with the distance. However, Carl and Kühn (2007) argued that “observations close to each other geographically are more likely to be similar than those far away from each other”. In other word, in general, the weights are inversely related with the distance, i.e. we should assign greater weights to the observations close to the given observations and smaller weights to the observations that are far away from the given observations. On the other hand, we still want to keep the local adaptive nature of the spatial weight matrix $W^*$, that is why $W^*$ outperforms the other spatial weight matrix. Thus, in this study, we use inverse-distance to weight the Getis–Ord $G_i^*(d)$ in order to obtain a new spatial weight matrix. The modified spatial weight matrix $W^{**}$ is defined as:

1. if $d_c = 0$, $w_{ij}^{**} = 0$, for all $j$;
2. if $d_c = d_i$, $w_{ij}^{**} = 1$ for all $j$, where $d_{ij} = d_c$; $w_{ij}^{**} = 0$, otherwise;
3. if $d_c > d_i$, $w_{ij}^{**} = \left(\frac{G_i^{*}(d_c)}{d_c^k} - \frac{G_i^{*}(d_{ij})}{d_{ij}^k}\right) - \left(\frac{G_i^{*}(d_c)}{d_c^k} - \frac{G_i^{*}(0)}{0^k}\right)$, for all $j$, where $d_{ij} \leq d_c$; $w_{ij}^{**} = 0$, otherwise;

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

The Local Statistics Model (LSM) is used in the simulation study as follow:

$$Y = \alpha + \rho \cdot Y + \beta X + \epsilon,$$

(2-4)

where $Y$ is the response variable generated by simulation; $\rho$ represents the dependence structure of the variable $Y$; $\beta$ equates the effect on the observations that are not correlated
with any of their neighbors; and $W$ is a spatial weight matrix ($W^*$ or $W^{**}$). Same as Getis and Aldstadt (2004), a dummy variable $X$, that takes on the value one for all observations having no dependence structure and zero otherwise, is created to compensate the zero-rows effects in $W$. The parameters are estimated using maximum-likelihood methods.

**Simulation Design**

For the simulation study, we use the same design as Getis and Aldstadt (2004) except that we change the radius of 2-cluster to 6 instead of 8 in order to avoid overlap of cells (Table 2.1). Three types of 30 by 30 raster data set are simulated and each type is simulated 25 times. The first type, a random normal, represents the situation that there is no spatial autocorrelation among the values placed in the cells. The second type displays a pattern of 2 clusters indicating the spatial structure with 2 opposite patches patterns. The third type displays a pattern of 6 clusters indicating the spatial structure with randomly combination of multi-patches patterns. All the values put in the cells are generated from a standard normal distribution. The 50 replications with cluster pattern are supposed to be sufficient to represent a wide variety of spatial structures (Getis and Aldstadt 2004). Figures 2.1 through 2.3 respectively show one realization of the random normal pattern, 2-cluster pattern, and 6-cluster pattern. Figures 2.4 through 2.6 respectively show the spatial distribution of the critical distances for the data sets shown in Figures 2.1 through 2.3. For the $i^{th}$ observations, the critical distance is calculated as following: First, for each given distance $d$, we find the nearest neighbors of the $i^{th}$ observations within distance $d$; Second, we calculated the $G_i^*(d)$; At last, the critical distance $d_c$ is the value $d$ such that $G_i^*(d)$ is the maximum and $G_i^*(d)$ is monastically increasing in the interval $(0, d_c)$.
<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random N=25</td>
<td>Random placement of values sampled from the normal distribution with mean 0, and standard deviation 1; 25 simulations</td>
</tr>
<tr>
<td>2-Cluster N=25</td>
<td>1 cluster of high values at (10, 10) with radius 6 and 1 cluster of low values at (20, 20) with radius 6—values from the normal distribution with mean 0, and standard deviation 1; 25 simulations. The highest values from the random generation were placed randomly in the high value cluster, while the lowest were placed randomly in the cluster of low values. The remaining values, those in the middle of the distribution, were placed randomly outside the clusters.</td>
</tr>
<tr>
<td>6-Cluster N=25</td>
<td>6 randomly placed clusters, 3 of high values and 3 of low values with radii 2, 4, and 6 respectively; values are sampled from the normal distribution with mean 0, and standard deviation 1; 25 simulations. As in the two-cluster case the highest values were placed randomly, but this time in the three high value clusters. The low values were placed randomly in the three low value clusters. The remaining middle values were placed randomly outside the clusters.</td>
</tr>
</tbody>
</table>

Figure 2.1 Random Data Set. 900 values are generated from the standard normal distribution and randomly assigned to the cells of the 30 by 30 grid.
Figure 2.2  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).

Figure 2.3  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 (14, 27), (27, 14), (8, 22), (22, 8), (10, 10), (20, 20).
Figure 2.4  Critical distance ($d_c$) for the random pattern in Figures 2.1. Distances are based on nearest neighbors.

Figure 2.5  Critical distance ($d_c$) for the 2-Cluster pattern in Figures 2.2. Distances are based on nearest neighbors.
Evaluation Criteria

Per Getis and Aldstadt (2004), the Akaike Information Criterion (AIC), autocorrelation coefficient $\rho$, and Moran’s I of residuals are used to evaluate the model performance.

AIC, developed by H. Akaike (1974), is a measure of the relative goodness of fit of a statistical model. AIC is defined as: $AIC = 2 * k - 2 * Ln(L)$, where $k$ is the number of parameters in the estimated model and $L$ is the maximized value of the likelihood function for the estimated model. 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) argued that “the autocorrelation coefficient gives an interpretation for the possible association between $W_y$ and $y$”. If $\rho = 1$, it means that the
spatial weight matrix $W$ is a good representation of the spatial autocorrelation among data; otherwise, if $\rho$ is close to 0, it means that the spatial weight matrix $W$ 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 defined as:

$$I = \frac{N^* \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (e_i - \bar{e})(e_j - \bar{e})}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} (e_i - \bar{e})^2}$$

(2-5)

where $N$ is the number of cells, $e$ is the residuals vector, $w_{ij}$ is a spatial weight between the $i$th observation and $j$th observation. If the $W$ matrix completely account for all of the spatial variation in $y$, the residuals should be spatially random. Moran’s I of the residuals is computed using the same spatial weight matrix $W$ that is used to estimate the corresponding model.

**Results and Discussion**

Table 2.2 displays the results of simulation. LSM 1 is the LSM model using the spatial weight matrix $W^*$ and LSM 2 is the LSM model using the spatial weight matrix $W^{**}$. The simulation results show that: in all the cases, the value of autocorrelation coefficient, $\rho$, are significantly different from 0 (p-values are less than 0.0000 respectively) for both models.

As for Moran’s I, in general, the range of Moran’s I is in the interval (-1, 1). But, in the 2- and 6-cluster cases, both values of Moran’s I are far greater than 1. This may attribute to that the number $N$ in (2-5) does not represent the true number of cells should
be considered since some of observations have no spatial autocorrelation with its neighbors and the corresponding rows in the modified spatial weight matrices are zero.

In random case, the mean AIC values and the variation of AIC are almost the same for both LSM models; in the 2-cluster case, the mean AIC value of LSM 2 is not significantly different from that of LSM 1; in 6-cluster case, the mean AIC value of LSM 2 is significantly smaller than that of LSM 1 at 10% significance level (the p-values is 0.0755). However, the estimated mean values of autocorrelation coefficient, $\rho$, and the Moran’s I value of residuals have small but significant difference for both LSM models in both the 2- and 6-cluster cases. Thus, we conclude that the spatial weight matrix $W^{**}$ performs better than the spatial weight matrix $W^*$ according to AIC rule when the spatial autocorrelation occurred in the data.

<table>
<thead>
<tr>
<th></th>
<th>AIC</th>
<th>Estimated $\rho$</th>
<th>Moran’s I of residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LSM 1</td>
<td>LSM 2</td>
<td>LSM 1</td>
</tr>
<tr>
<td><strong>Random</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2486.6</td>
<td>2486.8</td>
<td>0.78</td>
</tr>
<tr>
<td>Max</td>
<td>2613.6</td>
<td>2613.3</td>
<td>0.88</td>
</tr>
<tr>
<td>Min</td>
<td>2404.3</td>
<td>2404.5</td>
<td>0.67</td>
</tr>
<tr>
<td>SD</td>
<td>56.2</td>
<td>56.2</td>
<td>0.05</td>
</tr>
<tr>
<td><strong>2-Cluster</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1677.9</td>
<td>1663.0</td>
<td>1.17</td>
</tr>
<tr>
<td>Max</td>
<td>1796.1</td>
<td>1789.5</td>
<td>1.20</td>
</tr>
<tr>
<td>Min</td>
<td>1550.5</td>
<td>1533.2</td>
<td>1.13</td>
</tr>
<tr>
<td>SD</td>
<td>64.5</td>
<td>66.8</td>
<td>0.02</td>
</tr>
<tr>
<td><strong>6-Cluster</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>1420.2</td>
<td>1400.1</td>
<td>1.18</td>
</tr>
<tr>
<td>Max</td>
<td>1492.8</td>
<td>1481.9</td>
<td>1.20</td>
</tr>
<tr>
<td>Min</td>
<td>1318.8</td>
<td>1295.1</td>
<td>1.14</td>
</tr>
<tr>
<td>SD</td>
<td>47.7</td>
<td>49.8</td>
<td>0.02</td>
</tr>
</tbody>
</table>
According to the definition of $w^{**}_{ij}$,

$$w^{**}_{ij} = \frac{G_i^*(d_{ij})/d_{ij} - G_i^*(d_c)/d_c}{G_i^*(d_{ij})/d_{ij} - G_i^*(0)/d_c} \approx \frac{G_i^*(d_{ij}) - G_i^*(d_c) \times d_{ij}/d_c}{G_i^*(d_{ij}) - G_i^*(0) \times d_{ij}}$$

(2-6)

In general, if $d_{ij}$ is close to $d_c$, in other word, $d_c / d_{ij}$ is close to 1, then $w^{**}_{ij}$ is greater than $w^*_{ij}$; if $d_{ij}$ is far smaller than $d_c$, in other word, $d_c / d_{ij}$ is large enough, then is smaller than $w^*_{ij}$. Thus, the modification of $W^*$ adjusts weights by assigning greater weight to the observations which is far away from $i^{th}$ observations and assigning smaller weight to the observations which is close to $i^{th}$ observations. The better performance of the spatial weight matrix may attribute to that $W^*$ over-weighted the observations that are close to $i^{th}$ observations and under-weighted the observations that are far away from $i^{th}$ observations.

**Summary**

In this study, we extend the spatial weight matrix defined in Getis and Aldstadt (2004) to a more general case. The modified spatial weight matrix performs better than that of Getis and Aldstadt (2004) based on AIC since it takes account of the distance between observations to adjust the weights of observations.

Notice that the spatial weight matrix $W^*$ is a special case of the spatial weight matrix $W^{**}$ as $k = 0$. This study extends the spatial weight matrix $W^*$ to a more general case and still keeps its local adaptive property. In this study, $k$ is set to one for simplicity. However, we may choose an optimal $k$, which minimize AIC, through trial-error method for the data. On the other hand, the other local spatial statistics, such as local Moran’s I and Geary’s C may be used to construct spatial weight matrices.
CHAPTER III

DISPERSAL OF MULTIFLORA ROSE IN THE UPPER MIDWEST

Introduction

In recent years, the presence and impact of non-native invasive plants (NNIPs) has become one of the top concerns in natural resources management. Some NNIPs can displace native plants, negatively affect ecosystems, threaten native biodiversity, and lower the value of local ecosystems due to their competitive advantage compared with native plants (Macdonald 1994 and Moser et al. 2009). In the U.S., the estimated economic loss of agricultural crops due to alien invasive weeds was more than $24 billion per year (Pimentel et al., 2005).

Theoharides and Dukes (2007) summarized and qualitatively classified the invasive process of NNIPs into four stages: introduction, colonization, establishment and spread. In introduction stage, NNIPs are transported from its native regions to a new region via long distance movements such as the global commerce and travel. In the colonization stage, NNIPs survive and achieve positive growth rates at low densities in the infested region. In the establishment stage, NNIPs develop self-sustaining and expanding populations. In the spread stage, NNIPs disperse within a region over significantly time periods. However, this classification is not conclusive. Davis (2009) summarized the literatures about the study of invasion process. The invasion process may be classified into two stages (Davis 2009), three stages (Williamson 1996, Radosevich 2007), or up to six stages (Henderson et al. 2006). These studies have the same
characteristics: these classifications are qualitative and based on the life-history traits of the species, the genetic traits of the species, or environmental factors etc.

On the other hand, in forested regions (Hereafter, discussion of NNIPs will pertain to forested ecosystems.), the factors that affected the invasion of NNIPs include temperature, site quality, stands size, forest fragmentation, distance to roads, and forest density (e.g., here represented by the percentage of total county area in forest) (Moser et al. 2009). Each factor may play a role in one or more stages. The management of NNIPs depends on the stages of invasion since the cost dramatically increases as the populations of NNIPs expand (Hobbs and Humphries 1995). It is prohibitively expensive to remove most of the NNIPs from the invaded region after they are well established, so efforts to control invasive plants should focus on the establishment or earlier stages (Webster et al. 2006). Thus, the early detection of NNIPs is very important for control of NNIPs.

In this study, we focus on the evaluation of multiflora rose (*Rosa Multiflora*, Thunb. *ex* Murr.) (MFR) since it has highest occupancy rate on the FIA plots in seven Upper Midwest States. MFR is a thorny, perennial shrub. MFR is native to eastern Asia and was introduced to North America in 1866 as rootstock for ornamental roses. Later, it was used for erosion control and "living fences" to confine livestock from the 1930s to the 1950s (Doll 2006) and as crash barriers to reduce headlight glare in the medians of highways (BHWP 1997). MFR can exclude native ground flora, suppress tree regeneration, and has been designated as a noxious weed in many states such as Wisconsin, Missouri, Pennsylvania, Virginia, Iowa, and West Virginia (Munger 2002, Denight el al. 2008). MFR has adapted to many North America habitats (Doll 2006) and presents widely across the U.S. In fact, 38 state in the contiguous US report the presence of MFR (Denight el al. 2008). MFR grows in roadsides, old fields, pastures, prairies,
savannas, open woodlands, forest edges, and disturbed dense forests. MFR is most productive in sunny areas with well-drained moist uplands and bottomlands; it endures shade, sun, damp and/or dry conditions, but does not grow well in standing water or in extremely dry areas (Munger 2002; BHWP 1997). MFR cannot tolerate extreme cold temperatures (Denight et al 2008). MFR is extremely prolific. Its seeds are primarily dispersed by birds and may remain viable in the soil for many years. Control methods include mechanical and chemical methods that require repeated application for success, making control very expensive (Munger 2002).

Richardson and Pyšek (2006) summarized four factors that influence invasion success: human disturbance, competitive release, resource availability, and propagule pressure. Human disturbance plays an important role. Forest fragmentation, as one of the human disturbance activities, is the process of breaking up large patches of a forest into smaller pieces. Both human activities and natural events may result in forest fragmentation, such as road development in forest, land use change, wildfire. Forest fragmentation provides openings into the forest, and frequently facilitates invasion of NNIPs.

Roads provide not only habitat but also corridors of dispersal in the spread and growth of NNIPs (Mortensen et al. 2009). Thus, the density of roads per unit area (say, a county) or the distance to the nearest road, as a proxy of human disturbance, is one of the important factors that contribute to the spread of NNIPs. Road development may result in increasing forest fragmentation and which, in turn, may significantly alter landscape structure (Saunders et al. 2002). As road density increases, a higher proportion of the landscape becomes roadside habitat, which tends to be highly invaded by invasive plants (Gelbard and Belnap 2003). Watkins et al. (2003) documented that prevalence of NNIPs
is negatively associated with the distance to road and “exotic species are most prevalent within 15 m of roads, occurring infrequently in the interior forest”. Von der Lippe and Kowarik (2007) found that long-distance dispersal of seeds of NNIPs by vehicles is a routine rather than an occasional mechanism, and dispersal of plants by vehicles will accelerate plant invasions. Moser et al (2009) found that the presence of MFR is significantly and negatively related with distance to road at the plot level. Therefore, in this study, we expect to see a positive relationship between the presence of MFR and the density of roads, in other word, the greater density of roads favors the invasion of MFR.

The objectives of this study are: 1) develop a procedure to quantify the invasion stages at county level; 2) identify the driving factors that influence the dispersal of MFR in the seven Upper Midwest at broader scale – county level; and 3) compare the performance of the spatial weight matrices defined in Chapter II and answer the question: do the spatial weight matrices affect the choice the set of factors that used to explain the ecological phenomena?

**Study Area**

The Midwest study area is comprised of seven states: Indiana, Illinois, Iowa, Missouri, Michigan, Wisconsin, and Minnesota. At the nexus of several ecoregions, this area is characterized by diverse vegetation compositions and structures. Northern Minnesota, northern Wisconsin, northern Michigan, and southern Missouri are the most heavily forested areas (31.2% of total land in Michigan is covered by forest; 38.5% in Wisconsin, 29.3% in Minnesota and 32.8% in Missouri) (Table 3.1). The middle of this area, once covered mostly by prairie during pre-European settlement times, is currently a mosaic of agricultural lands embedded with urban areas (8% of total land in Iowa is
covered by forest; 12.2% in Illinois and 20.3% in Indiana) (Table 3.1). Extensive human activities and fertile soil in these states favor the establishment of NNIPs. In this study area, the primary forest type groups (Table 3.2) are oak/hickory (37.8%), maple/aspen/beech/birch (29.6%), and elm/ash/cottonwood (10.9%). In 2000, oak-hickory is dominant in Missouri, Illinois, and Iowa; maple/beech/birch is dominant in Michigan, Wisconsin, and Indiana; and aspen/birch is dominant in Minnesota (Potts et al. 2004).

Data

In this study, U.S. Forest Service, Northern Research Station Forest Inventory and Analysis (FIA) data from the Upper Midwest states for the period 2005-2006 are used. Phase 2 (tree inventory) data are collected on the standard FIA plot grid (1 plot per 2400 ha). Each phase 2 plot consists of 4 subplots with a radius of 7.3 m. Associated with this overstory inventory data is sampling for 25 invasive plant species of interest. These 25 species are categorized as grasses, vines, herbaceous species, or woody species, such as MFR. In this dataset, two variables, presence and cover rate (see Table 3.3 for cover code and the corresponding range of percent cover), are used to describe the presence and abundance of each invasive plant in each plot. For each invasive plant, the values of the presence are 0 (absence) or 1 (presence); the values of cover rate are recorded as the middle value of the range of percent cover for each cover code (Table 3.3). In total, 8632 phase 2 forested plots are assessed for MFR, where 1320 plots (15.2%) are invaded by MFR. 594 counties out of 649 counties in the Upper Midwest have FIA plots.
Table 3.1  Forested land in seven Upper Midwest States

<table>
<thead>
<tr>
<th>State</th>
<th>Forested Land (in Acres)</th>
<th>Percentage of forested land</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illinois</td>
<td>4,525,261</td>
<td>12.2%</td>
</tr>
<tr>
<td>Indiana</td>
<td>4,726,340</td>
<td>20.3%</td>
</tr>
<tr>
<td>Iowa</td>
<td>2,878,942</td>
<td>8.0%</td>
</tr>
<tr>
<td>Michigan</td>
<td>19,338,553</td>
<td>31.2%</td>
</tr>
<tr>
<td>Minnesota</td>
<td>16,300,691</td>
<td>29.3%</td>
</tr>
<tr>
<td>Missouri</td>
<td>14,646,319</td>
<td>32.8%</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>16,119,339</td>
<td>38.5%</td>
</tr>
</tbody>
</table>

(Source: Forest Inventory and Analysis Data 2005)

Table 3.2  Area of sampled forested land by States and Forest type group (in km$^2$)

<table>
<thead>
<tr>
<th>Forest-type group</th>
<th>State</th>
<th>Conifer</th>
<th>Oak / pine</th>
<th>Oak / hickory</th>
<th>Oak / tupelo</th>
<th>Elm / ash / cottonwood</th>
<th>Maple / Aspen / birch</th>
<th>Other hardwoods</th>
<th>Exotic hardwoods</th>
<th>Nonstocked</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illinoi</td>
<td>360</td>
<td>278</td>
<td>13610</td>
<td>218</td>
<td>4713</td>
<td>298</td>
<td>29</td>
<td>6</td>
<td>170</td>
<td>19683</td>
<td></td>
</tr>
<tr>
<td>Indiana</td>
<td>395</td>
<td>525</td>
<td>14157</td>
<td>263</td>
<td>2413</td>
<td>1349</td>
<td>57</td>
<td>3</td>
<td>163</td>
<td>19324</td>
<td></td>
</tr>
<tr>
<td>Iowa</td>
<td>110</td>
<td>363</td>
<td>8296</td>
<td>--</td>
<td>2785</td>
<td>280</td>
<td>92</td>
<td>49</td>
<td>272</td>
<td>12246</td>
<td></td>
</tr>
<tr>
<td>Michigan</td>
<td>19037</td>
<td>2386</td>
<td>13254</td>
<td>49</td>
<td>8318</td>
<td>36621</td>
<td>508</td>
<td>48</td>
<td>730</td>
<td>80951</td>
<td></td>
</tr>
<tr>
<td>Minnesota</td>
<td>20127</td>
<td>1275</td>
<td>8543</td>
<td>--</td>
<td>6660</td>
<td>31515</td>
<td>959</td>
<td>35</td>
<td>859</td>
<td>69974</td>
<td></td>
</tr>
<tr>
<td>Missouri</td>
<td>2544</td>
<td>3729</td>
<td>50932</td>
<td>272</td>
<td>4551</td>
<td>311</td>
<td>105</td>
<td>14</td>
<td>243</td>
<td>62701</td>
<td></td>
</tr>
<tr>
<td>Wisconsin</td>
<td>12393</td>
<td>2437</td>
<td>17220</td>
<td>--</td>
<td>6980</td>
<td>28142</td>
<td>291</td>
<td>20</td>
<td>796</td>
<td>68280</td>
<td></td>
</tr>
<tr>
<td>Totals:</td>
<td>54966</td>
<td>10992</td>
<td>126013</td>
<td>803</td>
<td>36420</td>
<td>98515</td>
<td>2042</td>
<td>175</td>
<td>3233</td>
<td>333160</td>
<td></td>
</tr>
</tbody>
</table>

Percentage: 16.5% 3.3% 37.8% 0.2% 10.9% 0.6% 0.1% 1.0%

(Source: Forest Inventory and Analysis Data 2006-2010)

Table 3.3  Cover codes and ranges of percent cover of non-native invasive plants used in recording invasive species’ presence, FIA plots, 2005-2006

<table>
<thead>
<tr>
<th>Cover code</th>
<th>Range of percent cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt; 1 percent, trace</td>
</tr>
<tr>
<td>2</td>
<td>1 to 5 percent</td>
</tr>
<tr>
<td>3</td>
<td>6 to 10 percent</td>
</tr>
<tr>
<td>4</td>
<td>11 to 25 percent</td>
</tr>
<tr>
<td>5</td>
<td>26 to 50 percent</td>
</tr>
<tr>
<td>6</td>
<td>51 to 75 percent</td>
</tr>
<tr>
<td>7</td>
<td>76 to 100 percent</td>
</tr>
</tbody>
</table>
Methodology

The presence probability of MFR

In the data, on average, there are 15 plots in each county. We notice that 59 counties only have 1 FIA plot and 22 of which were infested by the MFR (Figure3.1). Thus, the presence probability is 100% in these counties if we define the presence probability as the ratio between the number of the MFR-presence plots and the total number of the FIA plots in the county. On the other hand, 65 counties have more than 30 FIA plots including 3 counties which have only one MFR-presence FIA plot. We observe the fact that the number of plots in each county does affect the estimation of the presence probability. In order to overcome the bias due to the sample size, we use the following formula to define the presence probability:

\[ p_i = \frac{\sum_{j\in \eta_i} s_j}{\sum_{j\in \eta_i} n_j} \]  

(3-1)

where \( s_j \) is the number of the MFR-presence plots in the county \( j \), \( n_j \) is the total plots in the county \( j \), \( \eta_i \) is the set of counties that share boundary with the county \( i \), including the county \( i \).
Moran’s I and Geary’s C (Waller and Gotway 2004)

To investigate the spatial correlation of the presence probability of MFR, Moran's I and Geary's C are calculated as:

\[
Moran's \ I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} (Y_i - \bar{Y})^2}
\]  \hspace{1cm} (3-2)

and

\[
Geary's \ C = \frac{(N - 1) \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (Y_i - Y_j)^2}{2 \sum_{i=1}^{N} (Y_i - \bar{Y})^2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}}
\]  \hspace{1cm} (3-3)

where \( N \) is the number of counties; \( Y \) is the presence probability at each county; \( w_{ij} \) is a binary spatial weights matrix.
In general, the values of Moran’s I range from -1 to 1. The negative value of Moran’s I indicates negative spatial autocorrelation; the positive value of Moran’s I indicates positive spatial autocorrelation; and a zero value indicates no spatial autocorrelation. The range of the value of Geary’s C is (0, 2). If the value of Geary’s C is between 0 and 1, positive spatial autocorrelation is indicated; if the value of Geary’s C is between 1 and 2, negative spatial autocorrelation is indicated; and the value of Geary’s C equals to 1 indicates no spatial autocorrelation. Geary's C is inversely related to Moran's I and is more sensitive to local spatial autocorrelation.

**Classification and Regression Tree Method (CART)**

Classification and regression tree (CART) methods (Breiman et al. 1984) are used to classify 649 counties in the Upper Midwest into different categories based on the estimated presence probability and the estimated cover rate of MFR. CART is a non-parametric statistical method and has no assumptions about the distribution of dependent variables. It can be used for predicting continuous dependent variables (the resultant is called as regression tree) and categorical dependent variables (the resultant is called as classification tree). It also can be used to handle a set of continuous or categorical independent variables or a mixture of them. CART is a recursive partitioning method that partitions a heterogeneous population into a set of relatively homogeneous populations. The result of CART is the graph of the classification (or regression) tree. The tree consists of parent nodes, child nodes, and terminal nodes. For the binary case, a parent nodes split to obtain two child nodes. The terminal nodes are the nodes that do not have child nodes in the tree. CART is composed of two processes: growing the tree and optimizing the tree. In order to grow a tree, three criteria are needed: 1) the selection of
splits; 2) a rule to stop splits; 3) a rule for assigning a terminal node to a class (Breiman et al. 1984). There are a few techniques used to select the splits. But the most commonly used is the Gini improvement measure. For the binary case, the Gini Index is defined as: $Gini(t) = 1 - p_0^2 - p_1^2$, where $p_i$ is the relative proportion of category $i$ in node $t$. The Gini Index of the tree is defined as:

$$
Gini_{tree} = \frac{1}{N} \sum_{i=1}^{n} N_i * Gini_i 
$$

where $n$ is the number of the terminal nodes; $N_i$ is the size of the terminal node $i$; $Gini_i$ is the Gini Index at terminal node $i$. And the Gini improvement measure, which is used to choose the best split, is defined as:

$$
\Delta I(s,t) = Gini(t) - p_L * Gini(t_L) - p_R * Gini(t_R)
$$

where $t_L$ and $t_R$ are the left and right child nodes of a node $t$; $p_L$ and $p_R$ are the proportion of samples in node $t$ split into left and right child nodes. Without stopping rules often results in an over-fitted tree, which is hard to interpret.

To optimizing the tree means to prune the over-fitted tree. Maindonald and Braun (2007) described the details of the pruning method based on the rpart package (Therneau and Atkinson 1997) in R language. The rpart package automatically performs cross-validation and output the cross-validation relative error rate (the column headed xerror in the output), which estimates the expected error rate for use of the prediction tree with new data. The cross-validation relative error rate gives an assessment of the performance of CART, i.e. the change in prediction error with changing tree size. The cross-validation relative error rate decreases as the number of splits increase, which is one less than the tree size (the number of the terminal nodes), then increases as the number of splits is
large enough. Therefore, the optimal tree is the tree with the number of splits that minimize the cross-validation relative error rate (Maindonald and Braun 2007).

**Principle Component Analysis (PCA)**

PCA (Pearson 1901) is a widely used multivariate technique for dimension reducing. PCA uses the orthogonal transformation to find a set of fewer and uncorrelated variables (principle components) to represent the original set of variables without much loss of information or accounting for as much of the variation of the original variables as possible. PCA can be done by eigenvalue decomposition of the covariance matrix of the original variables.

In this study, we use PCA to obtain one principle component from the estimated presence probability and the estimated cover rate. First, the covariance matrix of two variables is computed after they are respectively centered with their mean. Second, we calculate the two eigenvalue of the covariance matrix, denoted as $\lambda_1$ and $\lambda_2$ (assume $\lambda_1 > \lambda_2$), and the corresponding eigenvectors. Third, choose the eigenvalue $\lambda_1$ since $\lambda_1 > \lambda_2$, and its corresponding eigenvector if $\lambda_1 / (\lambda_1+\lambda_2) > 0.9$, otherwise, we do not transform the variables. At last, the two components of the chosen eigenvector are used to weight the two estimated variables in order to obtain the principle component. And the variation explained by the principle component is: $\lambda_1 / (\lambda_1+\lambda_2)$.

**Classification of invasion stages**

The following classification procedures are based on the conceptual diagram of the occupied area versus time (Radosevich 2007) based on the work Cousens and Mortimer (1995). In this diagram, the occupied area is an increasing function of time. Radosevich (2007) classifies the invasion process into three invasion stages: introduction,
colonization, and naturalization. Based on this classification, the curve in the diagram is classified into three segments and they are corresponding to the three stages. In the introduction stage, the occupied area slowly increases as time; in the colonization stage, the occupied area quickly increases as time; and, in the naturalization stage, the occupied area slowly increases and gradually reaches a stable limit or carrying capacity. In each of these stages, the corresponding management strategies are different. When the abundance of the NNIPs is very low, the corresponding management strategies are quarantine or eradication priority; when the abundance of the NNIPs increases very fast such as that NNIPs is in the colonization stage, the corresponding management strategy is control priority; and the abundance of the NNIPs is very high (in the naturalization stage), the effective control becomes impossible without massive resource inputs. In this study, we try to develop the procedures to classify the invasion stages based on the thought behind the conceptual diagram in Radosevich (2007). However, the classification of invasion stages is not necessary to be the same as that in Radosevich (2007).

**Univariate Method**

In this case, we only consider the information implied in the data of MFR presence/absence. We use the following procedure (Figure 3.2) to classify the counties into different invasion stages:

First, based on the estimated presence probability, we calculate the cumulative distribution function (CDF) of the estimated presence probability (pp) at county level. The interval [0, 1] is divided into n small intervals with length h (n * h = 1). For each i ≤ n, we count the number of counties that the estimated presence probability is not greater than i/n. Then the value of the CDF for the given value i/n is the ratio of the number of
counties and the total number of counties. In this study, we set $h = 0.001$ and $n = 1000$, which is greater than the total number of counties 649 and the computation time is acceptable.

Second, we use the CART method to segment the CDF of the estimated presence probability. The calculated CDF of presence probability serve as the response variable and the value $i/n$, $i$ from 0 to $n$, serve as the independent variable. To choose the number of the tree size, which is defined as the number of the terminal nodes, the X–relative errors (cross-validation relative errors) are plotted against the number of tree sizes. Then the cutoff points are obtained from the CART and the segments of CDF are marked as into different stages.

Third, for each stage, we pool all the plots in the counties that are classified into the same stage. Then we count the numbers of plots for each cover code and the total number of plots in this stage. At last, for the given stage and for each cover code, the proportion of plots is the ratio of the numbers of plots and the total number of plots in the given stages. These ratios are used to investigate the reasonability of the classification in the previous step.

Fourth, we plot the map of stages for the study area (Figure 3.8).
The classification procedure of invasion stages based on the estimated presence probability.

**Multivariate Method**

The dataset also includes the variables about the abundance of the MFR in the seven Upper Midwest states: the cover code, which is a category variable, representing the middle value of the range of percent cover in each plot corresponding to the cover code. Though the cover rate is very coarse and discontinuous, it provides information of abundance of MFR. Incorporating the information of abundance of MFR with the information of presence/absence of MFR may be helpful to accurately classify the invasion stages. The procedure is described as following:

First, we estimate the cover rate at county level using the same method that are used to estimated presence probability; the estimated cover rate is a continuous variable taking values between 0 and 1.
Second, we use the Principle Component Analysis (PCA) to obtain the primary component which extracts most of information implied in the two variables: the estimated presence probability and estimated cover rate.

Third, similarly with the univariate case, we use the CART to classify. In this case, the response variable is the primary component obtained in the previous step and the independent variables are the estimated presence probability and estimated cover rate.

Fourth, we plot the map of stages for the study area (Figure 3.13).

Figure 3.3   The classification procedure of invasion stages based on the estimated presence probability and cover rate.
**Other variables**

Forest fragmentation classification map layer (1km resolution), which is in GeoTIFF format, is downloaded from the National Atlas ([http://www.nationalatlas.gov](http://www.nationalatlas.gov)). There are 8 classes of fragmentation. We combine these forest cover type groups into 3 forest fragmentation type groups: *edge and perforated*, *interior*, and *patch*. Then we calculate the percentage area (area of each fragmentation class/ area of each county) for each forest fragmentation class in each county of the seven Upper Midwest States.

Forest cover type map layer, which is in GeoTIFF format, is downloaded from the National Atlas ([http://www.nationalatlas.gov](http://www.nationalatlas.gov)). A total 25 forest cover type groups are obtained from the Advanced Very High Resolution Radiometer (AVHRR) and Thematic Mapper (TM) imagery. We combine these forest cover type groups into 7 forest cover type groups: *conifer*, *oak/pine*, *oak/hickory*, *oak/gum/cypress*, *elm/ash-cottonwood*, *maple/Aspen/beech/birch*, and *non-forest*. Then we calculate the percentage area (area of each forest cover type/ area of each county) for each forest cover types in each county of the seven Upper Midwest States.

Road map layer (1km resolution), which is in Shapefile format, is downloaded from the National Atlas ([http://www.nationalatlas.gov](http://www.nationalatlas.gov)). We calculate the length of interstate highway and state highway for each county in the seven Upper Midwest States. Then the interstate highway density (*Rdens1*) and state highway density (*Rdens2*), with the unit km/km^2, were calculate as the ratio of the length of highway and the area of the county. Furthermore, we define other variables: the ratio of the interstate highway density (state highway density) and the county percent forest, denoted as *Interstate/CPF* and *State/CPF* respectively, which measures the intensity of human disturbance.
Simultaneous Autoregressive Model (SAR) (Waller and Gotway 2004)

We use the simultaneous autoregressive model (SAR) to identify the driving factors of the spreading of MFR with the estimated presence probability ($pp$) as the dependent variable. The SAR model is expressed as:

\[ PP = X\beta + \varepsilon \]
\[ \varepsilon = \rho W\varepsilon + \nu \]  

(3-6)

where $\beta$s are the parameters that need to be estimated, $\nu$ is the independent error vector and assumed to be normally distributed, $\rho$ is the simultaneous autoregressive error coefficient, and $W$ is the spatial weight matrix; $X$ is a set of independent variables including: longitude ($Lon$), latitude ($Lat$), forest fragmentation ($edge$ and $perforated$, $interior$), forest cover type groups ($conifer$, $oak/pine$, $oak/hickory$, $oak/gum/cypress$, $elm/ash/cottonwood$, $maple/aspen/beech/birch$), county percentage forest ($CPF$), highway density ($Rdens1$ and $Rdens2$), and the ratios: $Interstate/CPF$, $State/CPF$.

Software for graphs and computing

In this study, all statistical computation, analysis, and simulation are conducted under the R statistical environment (R Development Core Team 2011). The package `spdep` in R is used to fit the SAR model. The parameters are estimated using Maximum likelihood method. The `rpart` package in R is used to implement CART (Therneau and Atkinson 1997; John and John 2003). The `sp` and `maps` packages are used to draw graphics (Bivand et al. 2008).
Results and Discussions

Classification of invasion process

MFR is established in Missouri, Illinois, Indiana, Iowa, and southern Wisconsin and Michigan and more widely distributed in northern Missouri, southeast Iowa, central and northern Illinois and Indiana. MFR is not widely established in Minnesota, northern Wisconsin, and northern Michigan (Figure 3.4). There is a positive spatial autocorrelation among the MFR presence (Moral’s I = 0.93, Geary’s C = 0.08). The plot of kernel smoothed presence probability (Figure 3.4) indicates a strong invasive pattern in all directions, but particularly southward and eastward. Figure 3.6 shows the kernel smoothed cover rate. At county level, the estimated cover rate ranges from 0 to 21.96%. There is a positive spatial autocorrelation among the MFR abundance (Moral’s I = 0.85, Geary’s C = 0.15). The clusters of cover rate of MFR mainly locate in Illinois, Indiana and the south of Iowa.

Figures 3.7 through 3.10 display the result of classification of MFR invasion stages based on only the estimated presence probability. Figure 3.7 suggests that the optimal classification tree size should be 4 since the relative cross-validation error decreases as the classification tree size increases and the curve becomes flat when the classification tree size is greater than 4. Figure 3.8 suggests the following three break points for the estimated presence probabilities (which range from 0 to 1): 0.2955, 0.5105, and 0.6665. Thus, the estimated presence probabilities are classified into four intervals: (0, 0.2955), (0.2955, 0.5105), (0.5105, 0.6665), (0.6665, 1), respectively corresponding to the four stages of invasion. For each stage, we calculate the proportion of plots for each cover class. On average, the counties in “Spread” stage are infested by MFR earlier than that in “Introduction” stage. The plots in “Spread” stage are more possible to have a
greater coverage of MFR than that in “Introduction” stage. Thus, for the greater cover class, the proportion of plots in “Spread” stage should be greater than that in “Introduction” stage; for small cover class, we expect to see the reverse (Figure 3.9). Figure 3.10 displays the map of invasive stages based on the above classification. The statistics shows that 174 counties are not infected by MFR; in 107 counties, MFR is in the “Introduction” stage; in 125 counties, MFR is in the “Colonization” stage; in 127 counties, MFR is in the “Establishment” stage; and in 116 counties, MFR is in the “Spread” stage.

Figure 3.4 The distribution of FIA Phase 2 Plots without (green) and with the presence of MFR (red) in the Upper Midwest, 2005-2006
Figure 3.5  The smoothed presence probability of MFR in the Upper Midwest, 2005-2006.

Figure 3.6  The smoothed cover rate of MFR in the Upper Midwest, 2005-2006.
Figure 3.7  The plot of X relative error (relative cross-validation error) vs. the classification tree size (the number of terminal nodes).

Figure 3.8  Classification and regression tree partition of the estimated presence probability of MFR (pp-presence probability).
Figure 3.9  Proportion of plots vs. the midpoint value of cover class.

Figure 3.10  The map of invasive stages based on the estimated presence probability of MFR in the Upper Midwest, 2005-2006.

Figures 3.11 through 3.15 display the result of classification of MFR invasion stages based on both the estimated presence probability and estimated cover rate. The PCA results show that the primary component explained 91% variation among the estimated presence probability and the estimated cover rate. Figure 3.11 suggests that the optimal classification tree size should be 4 since the relative cross-validation error
decreases as the classification tree size increases and the decrease of relative error is very small when the classification tree size increases from 4 to 5. Figures 3.12 and 3.13 suggest the following classification: if the estimated presence probability is greater than 0 and not greater than 16.95%, then the county is classified into “Introduction” stage; if the estimated presence probability is greater than 16.95% and not greater than 39.67%, then the county is classified into “Colonization” stage; if the estimated presence probability is greater than 39.67% and the estimated cover rate is not greater than 8.14%, then the county is classified into “Establishment” stage; and if the estimated presence probability is greater than 39.67% and the estimated cover rate is greater than 8.14%, then the county is classified into “Spread” stage. Similarly for each stage, we calculate the proportion of plots for each cover class. Figure 3.13 shows that the estimated cover rate is non-linearly related with the estimated presence probability. As a whole, Figure 3.14 shows that the proportion of plots curve in each stage follows the expected as that in the case based on only the estimated presence probability. However, the “Spread” stage and “Introduction” stage Figure 3.15 displays the map of invasion stages based on the above classification. The statistics shows that 174 counties are not infected by MFR; in 69 counties, MFR is in the “Introduction” stage; in 83 counties, MFR is in the “Colonization” stage; in 208 counties, MFR is in the “Establishment” stage; and in 115 counties, MFR is in the “Spread” stage. Though the number of counties in “Spread” stage almost the same compared with the result that is solely based on the estimated presence probability, Figure 3.15 shows that the counties in the “Spread” stage are more clustering in the south Iowa, north and middle of Illinois, and north-west and middle of Indiana, compared with the Figure 3.10. On the other hand, the number of counties in “Introduction” stages decreases by 38 and the number of counties in “Colonization” stage decreases by 42; and
the number of counties in “Establishment” stage increases by 81. This may attribute to the fact that incorporating the abundance information of MFR corrects the inaccurate classification of invasion stages based on only the presence information of MFR.

Moody and Mack (1988) show that the infested area is a non-linear function of time if there are multiple foci of the invasion and the infested area is a linear function of time if there is single focus of the invasion. If we treat the estimated cover rate as a proxy of the infested area and the estimated presence probability as a proxy of the infested time, then the non-linear relationship between the estimated cover rate and the estimated presence probability (Figure 3.13) is consistent with the characteristics of MFR dispersal: a mixture of long-distance dispersal (by bird) and short-distance dispersal, and the history of MFR introduction, which was introduced into US in 1866 as rootstock for ornamental roses.

![Figure 3.11](image)

Figure 3.11  The plot of X relative error (relative cross-validation error) vs. the classification tree size (the number of terminal nodes).
Figure 3.12  Classification and regression tree partition of the estimated presence probability and estimated cover rate of MFR (pp-presence probability, cr-cover rate).

Figure 3.13  The plot of estimated cover rate vs. estimated presence probability and the result of classification and regression tree partition based on the estimated presence probability and estimated cover rate of MFR (pp-presence probability).
Figure 3.14  Proportion of plots vs. the midpoint value of cover class.

Figure 3.15  The map of invasive stages based on the estimated presence probability and estimated cover rate of MFR in the Upper Midwest, 2005-2006.

**SAR Model**

Table 3.4 displays the results of spatial autoregressive models (SAR). Model 1 and Model 2 are respectively the SAR model using the spatial weight matrix $W^*$ (based on Getis-Ord $G_i^*$) and $W^{**}$ (modified spatial weight matrix). The difference of estimated autocorrelation parameter $\rho$ is very small. The AIC value of Model 2 (-1265.40) is much smaller than that of Model 1 (-911.51) for the same set of independent variables. Since
the only difference between Model 1 and Model 2 is the spatial weight matrix, the
difference of AIC values in two models can be attributed to the difference in constructing
matrices. On the other hand, the Nagelkerke pseudo-R-squared of Model 2 (0.9048) is
smaller than that of Model 1 (0.8358). It means that the Model 2 explains more variations
of the MFR presence probability than that of Model 1. Thus, in this case, modified spatial
weight matrix $W^{**}$ performs better than $W^*$ since Model 2 has much smaller AIC value
and the Nagelkerke pseudo-R-squared. Model 2 are chosen to explain the variations
among the presence probability of MFR.

Notice that the forest cover type oak/hickory is significant in Model 1 but it is not
significant in Model 2. The ratio of the interstate road density and the county percentage
forest is not significant in Model 1 but it is significant in Model 2. The set of significant
variables of Model 2 is different from that of Model 1. Thus, we conclude that spatial
weight matrices do affect the choice of the factors used to explain the spread of MFR.

Table 3.4 Results of SAR models (part of non-significant variables is removed from
the models. This table only includes variables that are significant at 10%
significance level at least in one model.)

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Model 1</th>
<th>P-value</th>
<th>Model 2</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2.3191</td>
<td>&lt;0.0001</td>
<td>1.6653</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Longitude</td>
<td>0.0053</td>
<td>0.0012</td>
<td>0.0038</td>
<td>0.0023</td>
</tr>
<tr>
<td>Latitude</td>
<td>-0.0397</td>
<td>&lt;0.0001</td>
<td>-0.0286</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Oak/Gum/Cypress</td>
<td>-3.0239</td>
<td>0.0001</td>
<td>-1.1727</td>
<td>0.0533</td>
</tr>
<tr>
<td>Oak/hickory</td>
<td>-0.1485</td>
<td>0.0002</td>
<td>-0.0422</td>
<td>0.1802</td>
</tr>
<tr>
<td>Elm/Ash/Cottonwood</td>
<td>0.7781</td>
<td>0.0000</td>
<td>0.3496</td>
<td>0.0009</td>
</tr>
<tr>
<td>Interstate/CPF</td>
<td>0.0581</td>
<td>0.2251</td>
<td>0.1246</td>
<td>0.0005</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.9535</td>
<td>&lt;0.0001</td>
<td>0.9457</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>AIC</td>
<td>-911.51</td>
<td></td>
<td>-1265.40</td>
<td></td>
</tr>
<tr>
<td>Nagelkerke pseudo-R-squared</td>
<td>0.8358</td>
<td></td>
<td>0.9048</td>
<td></td>
</tr>
</tbody>
</table>
SAR (Table 3.5) shows that the Moran’s I is not significantly different from zero. It also shows that the presence probability of MFR is significantly and negatively associated with latitude and positively related with longitude. Combining with Figures 3.5 and 3.10, we conclude that from its central cores, MFR is invading northward into the forested lands in southern Wisconsin and Michigan, southward into the Ozark Highlands in southeast Missouri, and as well as toward the west. On the other hand, the negative relationship between the presence probability and the latitude, which is a coarse proxy of temperature, indicates that the northward spreading of MFR is limited by the extreme cold temperatures in the northern study area. This is consistent with the observation in Denight et al (2008).

The presence probability of MFR is also negatively associated with the forest cover type group oak/gum/cypress, which is bottomland hardwood forests and positively related with the forest cover type group elm/ash/cottonwood, which is upland hardwood forests. These relationships indicate that presence probability of MFR is more likely to invade the counties with higher proportion of upland forests and lower proportion of bottom land forests. These phenomena are consistent with the traits of the MFR habitat. MFR is most productive in sunny areas with well-drained moist uplands and bottomlands; it endures shade, sun, damp and/or dry conditions, but does not grow well in standing water or in extremely dry areas (Munger 2002; BHWP 1997). Thus, we conclude that the relative dryness and sun affects the invasibility of MFR.

The high density of highway and low county percent forest results in higher frequency of human disturbance. Thus, the ratio of the interstate highway density and the county percent forest (Interstate/CPF) measures the intensity of human disturbance. The presence probability of MFR is significant and positively associated with the ratio of the
interstate highway density and the county percent forest (Interstate/CPF). On the other hand, forest fragmentation and state road density, as the measures of human disturbance factors, are not significantly related to the presence probability of MFR. It may attribute to the relationship between the presence probability of MFR and the ratio of the interstate highway density and the county percent forest dominates the relationship between the presence probability of MFR and forest fragmentation and state road density. Thus, we conclude that increasing human disturbance favors the invasion of MFR in the study area.

Table 3.5 The final estimation of SAR models (This table only includes variables that are significant at 10% significance level at least in one model.)

<table>
<thead>
<tr>
<th>Model 2</th>
<th>Coefficients</th>
<th>Estimate</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.5675</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Longitude</td>
<td>0.0036</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td>Latitude</td>
<td>-0.0271</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
<tr>
<td>Oak/Gum/Cypress</td>
<td>-1.0032</td>
<td>0.0915</td>
<td></td>
</tr>
<tr>
<td>Elm/Ash/Cottonwood</td>
<td>0.3525</td>
<td>0.0008</td>
<td></td>
</tr>
<tr>
<td>Interstate/CPF</td>
<td>0.1319</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.9463</td>
<td>&lt;0.0001</td>
<td></td>
</tr>
</tbody>
</table>

Nagelkerke pseudo-R-squared 0.9046
Moran's I of Residuals -0.0296 0.7635

**Summary**

In this study, we compare the performance of two spatial weight matrices, $W^*$ and $W^{**}$, based on the local spatial statistics Getis-Ord $G_t^*$. We find that strong spatial autocorrelation exits among MFR presence and $W^{**}$ performs better than $W^*$. The choices of spatial weight matrix do affect the choice of the factors used to explain the spread of MFR.
We also develop procedures used to classify the invasion stages of MFR. The procedure based on the estimated presence probability and the estimated cover rate provides more accurate classification since it uses not only information of presence/absence of MFR but also the information of abundance of MFR. However, the coarse data of cover rate may influence the reliability and accuracy of the classification. Also, the management and control activities of MFR may affect the accuracy of classification.

The results support to classify the invasion process into four discrete stages. Different from the previous studies, these classifications procedures, based on the information of occupancy (presence/absence) and abundance of MFR, do not need the knowledge about the life-history of MFR, the genetic traits of MFR, or environmental factors etc.

The strong spatial autocorrelation of MFR presence and its aggressive spread southward and westward reflects the fact that human disturbances, climatic factors, and forest conditions play an important role. Even though MFR’s spread to northern states seems to be reduced due to cold temperatures, climate change and the increasing number of anthropogenic disturbances such as land clearing and urban development may increase the probability of infestation in the forests of the northern part of our study region.
CHAPTER IV
CONCLUSIONS

The primary objective of this study is to develop a spatial weight matrix based on local spatial statistics which can be used in the various spatial regression models. Its performance is examined through the simulation and an ecological application: the spread of invasive plants, MFR. The second objective of this study is to develop a procedure that can be used to classify the invasion stages of non-native invasive plants. On the other hand, we try to investigate the driving factors of the spread of MFR and the impacts of spatial weight matrix on the choice of the driving factors.

This study shows that the modified spatial weight matrix $W^{**}$ is a generalization of the spatial weight matrix $W^*$. The simulation study and the ecological example provide evidences that $W^{**}$ is outperformed than $w^*$. In the study of the invasive plants, the better performance of $W^{**}$ is quite clear. On the other hand, the spatial weight matrices do affect the choice of the set of factors that are used to explain the ecological phenomena.

The developed classification procedures of invasion stages are based on the information of occupancy and abundance of MFR. The results support to classify the invasion process into four discrete stages. Different from the other qualitative classification of invasion process, this study provide methods to quantify each invasion stages. The classification of the invasion stages becomes easy and feasible since the cutoffs values were obtained from the procedures. The map of invasion stages not only provides the visual insight on the infested areas, extent of MFR, and the direction of
MFR dispersal, but also helps the management to determine what method to be used (eradication or control) and how much resources should be allocated in each county. The non-linear relationship between the abundance and the occupancy of MFR is consistent with the observation that MFR primary dispersed by birds.

In this study, we document the strong spatial autocorrelation of MFR presence and find that human activities play important roles in the ecological processes. Anthropogenic disturbances may increase the probability of infestation in the forests of the northern part of our study region. Forest cover types also play important role in this ecological example.
REFERENCES


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


Geary, R. C., 1954, the Contiguity Ratio and Statistical Mapping, the Incorporated Statistician, 5 (3): 115–145


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


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


LeSage, J. P., Parent, O., 2007, Bayesian model averaging for spatial econometric models, Geographical Analysis, 39: 241-267


Moran, P. A. P., 1950, Notes on Continuous Stochastic Phenomena, Biometrika, 37, 17–33


Moser, W. K., Hansen, M. H., Nelson, M. D. [and others], 2009, Relationship of invasive groundcover plant presence to evidence of disturbance in the forests of the upper Midwest of the United States, Invasive Plants and forest ecosystems, 29-58, CRC Press


Webster, C. R., Jenkins, M. A., Jose, S., 2006, Woody invaders and the challenges they pose to forest ecosystems in the eastern United States, Journal of Forestry, 104(7): 366-374

APPENDIX A
CLASSIFICATION OF INVASION STAGES
Invasive Species Study

This study categorizes the invasion process into different invasion stages. The study is based on the Forest Inventory and Analysis dataset from U.S. Forest Service, Northern Research Station.

The study area is the Upper Midwest; the study period is 2005-2006.

Created by Weiming Yu, 10/21/2011

--- Loading Packages------------------------------------------------------------
library(maps)
library(mapproj)
library(maptools)
library(sp)
library(spdep)
library(spatstat)
library(stats)
library(rpart)

--- Lists
sn<-c('Illinois', 'Indiana', 'Iowa', 'Michigan', 'Minnesota', 'Missouri', 'Wisconsin')
ssn<-c('illinois', 'indiana', 'iowa', 'michigan', 'minnesota', 'missouri', 'wisconsin')
plot.stage=c("I", "II", "III", "IV")
level.val=c(0.01, 3, 8, 18, 38, 63, 87)
plot.col=c("blue", "green", "orange", "brown")
scc<-c("17", "18", "19", "26", "27", "29", "55")
drive<-c("F:")
path<-c("/2009_2_Fall_Coursework/FO_8313_Spatial_Stat/Final Proj/7states/")
dir<-paste(drive, path, sep="")

--- Self defined functions
read.csv.data<-function(directory, filename){
  read.csv(paste(directory, filename, sep=""))
}
presence.point<-function(data, color, pix=0.3){
  row.no=which(data[,3]<=0)
  data.presence=data[-row.no,]
  points(data.presence, cex=pix, pch=2, col=color)
}
spatial.plot<-function(data, SpatialPolygon, ID, title){
  spatialdataframe<-data.frame(data, row.names=as.character(ID))
  locMI.sppoly<-SpatialPolygonsDataFrame(SpatialPolygon, spatialdataframe, match.ID = TRUE)
x11()
  spplot(locMI.sppoly, col.regions=plot.col, col='blue', main=title)
}
order.dataframe<-function(data, var.num=2){
  for(i in var.num:1){ data<-data[order( data[,i] ), ] }
  return(data)
}
stages.plot<-function(IDs=ID.poly, n=stage.num){
  level=c("None", plot.stage[1:n])
  color=rep(0, n+1)
  color.list=c("white", plot.col)
  for (i in 1:(n+1) ) { color[i]=color.list[n+2-i] } 
  spatialdataframe<-data.frame(pstage, row.names=as.character(IDs))
  locMI.sppoly<-SpatialPolygonsDataFrame(shape, spatialdataframe, match.ID = TRUE)
  locMI.sppoly$pstage<-factor(locMI.sppoly$pstage, levels = rev(level))
x11()
  spplot(locMI.sppoly, col.regions=color, col="blue", scales=list(draw = TRUE), auto.key=TRUE, colorkey=TRUE)
}
#---- This function find the plots that the county is in the specific stage
stage.match=function(stage="Introduction", data=pstage, midvalue=cover.cat){
data.stage<-data.frame(county.id[,1], data)
data.stage<-data.stage[which(data==stage),]
nid<-c()
l1<-length(data.stage[,1])
for( i in 1: l1 ){
  nid<-c(nid, which(nnip$FIPS==data.stage[i, 1]))
}
data.midpoint<-cover.cat[nid]
data.midpoint<-data.midpoint[!is.na(data.midpoint)& data.midpoint>0]
t1<-table(data.midpoint)
propotion<-rep(0,7)
for( i in 1:7 ){
  propotion[i]=t1[i][1]
  if( is.na(propotion[i]) ) propotion[i]=0
}
return(propotion/sum(propotion))
}

#---- This function plot the proportion of plots curves
proportion.plot<-function( stagenum=stage.num){
  prop.mat=matrix(0,7,stagenum)
  for( i in 1:stagenum) { prop.mat[,i]=stage.match(plot.stage[i]) }
  barcol=gray(((stagenum-1):0)/(stagenum-1))
  x11()
  barplot(t(prop.mat),
    xaxt="n", xlab="midpoint of cover class", beside=TRUE,
    ylab="Proportion of plots", ylim=c(0, yrange), bty="n",
    col=plot.col[1:stage.num])
  axis(1, at=c(rep(1:7)-1)*5+3), level.val)
  legend("topright", plot.stage[1:stage.num], cex=1.5, bty="n",
    fill=barcol)
  box()
}

# The study of multiflora rose (MFR)

#---- read data
nnip<-read.csv.data(dir,"NNIP.csv")
FIPS<-nnip$State.code*1000+nnip$County.code
nnip<-data.frame( FIPS, nnip)
State<-nnip$State.code
Unit<-nnip$Unit.code
County<-nnip$County.code
Lon<-nnip$Lon.pub
Lat<-nnip$Lat.pub
mr<-nnip$Multiflora.rose
cover.cat<-nnip$Multiflora.rose.midpoint

#---- summary stats
pf=nnip$Percent.forest
row.pf=which(pf==1)
nobs=length(pf)
row.mr=which(mr==1)
num.pf=(nobs-length(row.pf))
num.mr=(length(mr)-1)
num.mr2=total.presents-length(row.pf2)
perc.presence=total.presents/nobs*100
name=c("Total observations", "mixed forest", "proportion", "Total presence", "percentage of presence",
  "mixed forest (presence)", "proportion (presence)"
stats=c(nobs, num.pf, num.pf/nobs*100, total.presence, perc.presence, num.pf2, num.pf2/total.presence*100)
num.summary=data.frame(name, round(stats) )

len<-length(mr)
unity<rep(1,len)
df.mr<-data.frame(Lon, Lat, mr, State, County, unity)
data.nnip<-data.frame(Lon, Lat, mr, cover.cat, State, County, unity)

#---- Plot the FIA plots

# the presence plot with county boundary
x11()
map1<-map('county', regions=ssn, col='blue')
points(Lon, Lat, cex=0.4, col='green')
presence.mr<-presence.point(df.mr, 'red')
box()

#---- Aggregate by county
a1<-aggregate(data.nnip, list(County, State), FUN=mean)
cfips<-as.character(a1$State*1000+a1$County)
Lon.lat<-a1$Lon
Lat.lat<-a1$Lat
mr.lat<-a1$mr
state.code<-a1$State
county.code<-a1$County
Total.plot<-a1$Total.plot
nnip.lat<-data.frame(cfips, state.code, county.code, mr.lat, cover.lat, Total.plot)

#---- Obtain windows
win1<locator()
win2<owin(poly=list(x=win1$x, y=win1$y))
shape<-readShapePoly(paste(dir, "7statesnew.shp", sep=""), IDvar="FIPS")
ID.poly<-as.character(county.id$Scid)
index<-match(county.id$Scid, shape@data$FIPS)
county.id<-data.frame(county.id, mr.ps, cover.ps, total.ps)

df.nnip<-data.frame(COUNTY, COUNTY.FIPS, COUNTY.NAME, STATE, STATE.FIPS, STATE.NAME)

cfips<-as.character(df.nnip$COUNTY.FIPS*1000+df.nnip$COUNTY)
Lon.x<-df.nnip$POINT_X
Lat.y<-df.nnip$POINT_Y

#---- Generate the neighbours and weights
mr.nb<-poly2nb(shape)
mr.w<-nb2listw(mr.nb, style="B")

#---- Calculate the presence probability
num.county<-length(county.id[,1])
prob.nb<-rep(0, num.county)
for( i in 1:num.county){
  t1<-0
  t2<-0
  index.list<-c(i, mr.nb[i])
  t1<-sum(county.id$mr.ps[ index.list ])
  t2<-sum(county.id$total.ps[ index.list ])
  if( t1<2 ) { prob.nb[i]<-0 } else { prob.nb[i]<-t1/t2 }
}
county.id<-data.frame(county.id, prob.nb)

#---- Calculate the cover rate
num.county<-length(county.id[,1])
cover.nb<-rep(0, num.county)
for( i in 1:num.county){
  t1<-0
  t2<-0
  index.list<-c(i, mr.nb[i])
  t1<-sum(county.id$cover.ps[ index.list ])
  t2<-sum(county.id$total.ps[ index.list ])
  if( t1<2 ) { cover.nb[i]<-0/100 } else { cover.nb[i]<-t1/t2/100 }
}
county.id<-data.frame(county.id, cover.nb)

#---- point pattern smoothing at county level
bandwidth<-0.14
x11()
prob.ppp<-ppp(county.id$Lon.x, county.id$Lat.y, window=win2, marks=county.id$prob.nb)
k.prob<-density(prob.ppp, kernel = "gaussian", bandwidth, weights=prob.ppp$marks)
plot(k.prob)
k1.prob<-density(prob.ppp, kernel = "gaussian", bandwidth)
plot(k1.prob, col="blue")
relative<-k.prob$v/k1.prob$v
k.prob$v<-relative
plot(k.prob, main=paste("Kernal Smoothing of Presence Probability (bandwidth="), bandwidth, ", sep=""))
map1<-map('county', regions=sn, col='black', add=T)
map1<-map('state', regions=sn, col='green', add=T)

#---- point pattern smoothing at county level
bandwidth<-0.125
x11()
cover.ppp<-ppp(county.id$Lon.x, county.id$Lat.y, window=win2, marks=county.id$cover.nb)
k.cover<-density(covar.ppp, kernel = "gaussian", bandwidth, weights=cover.ppp$marks)
plot(k.cover)
k1.cover<-density(covar.ppp, kernel = "gaussian", bandwidth)
plot(k1.cover, col="blue")
relative<-k.cover$v/k1.cover$v
k.cover$v<-relative
plot(k.cover, main=paste("Kernal Smoothing of Cover Rate (bandwidth="), bandwidth, ", sep=""))
map1<-map('county', regions=sn, col='black', add=T)
map1<-map('state', regions=sn, col='green', add=T)

#---- Generate the spatial statistics
moran.test(prob.nb, mr.w)
geary.test(prob.nb, mr.w)
moran.test(cover.rate, mr.w)
geary.test(cover.rate, mr.w)

# Classification - univariate case, based on the estimated presence probability

#---- Generate the CDF of the estimated presence probability
ns=1000
cdf.mm=rep(0,ns)
for( i in 1:ns){
  cdf.mm[i]=length( which(prob.nb<=i/ns))/649
}
pp=rep(1:ns)/ns
```r
rcontr<-rpart.control(minisplit=25, minbucket=2, cp=0.001)
cpf.split<-rpart(cdf.mr~pp, method="poisson", control=rcontr)
printcp(cpf.split)
x11()
plot(cpf.split$splitable[,"nsplit"]+1,cpf.split$splitable[,"xerror"],
     type="l", xlab="Tree Size", ylab="X relative error")

#----- Generate the prune plot
stage.num=4
split.num=stage.num-1
x11()

cp.strict<-cpf.split$splitable[stage.num, "CP"]
rcontr<-rpart.control(minisplit=25, minbucket=2, cp=cp.strict)
cpf.split<-rpart(cdf.mr~pp, method="poisson", control=rcontr)
printcp(cpf.split)
plot(cpf.split)
text(cpf.split)

#---- Find the splits location
splits=cpf.split$splits[1:split.num,]
splits=floor(splits[order(splits)]*1000)
splits
splits.loc=rep(0,3)
if (split.num>3 & split.num>0){
  splits=cbind(splits, rep(1000, 3-split.num))
}
splits.loc=splits

#---- Segament Plot
x11()
plot(rep(1:ns)/ns, cdf.mr, type="l",
      xlab="Presence Probability", ylab="CDF of Presence Probability")
new.sloc=c(1, splits.loc, 1000)
for (i in 1:4) {
  cdf.mrs=cdf.mr[new.sloc[i]:new.sloc[i+1]]
  lines(rep(new.sloc[i]:new.sloc[i+1])/ns, cdf.mrs, col=plot.col[i])
}
sp<-cpf.split$splits[1:split.num,]
sp<-sp[order(sp)]
for (i in 1:3) {
  lines(rep(splits.loc[i]:1000, 100), rep(1:100)/100, col="red")
text(splits.loc[i]/1000, 0.4, as.character(sp[i]))
}

#---- Classified each county into different stages and plot
perct=splits.loc/ns
pstage=rep(0, num.county)
for (i in 1:num.county) {
  if (prob.nb[i]==0) pstage[i]="None"
  if (prob.nb[i]<perct[1] & prob.nb[i]>=perct[0]) pstage[i]="Introduction"
  if (prob.nb[i]>=perct[3]) pstage[i]="Spread"
}

#---- The map of invasion stages
stages.plot()

length(which(pstage=="None"))
length(which(pstage=="Introduction"))
length(which(pstage=="Colonization"))
length(which(pstage=="Establishment"))
length(which(pstage=="Spread"))

#---- The proportion of plots
proportion.plot()
box()
```
# Classification - multivariate case, based on the estimated presence probability and cover rate

```r
cover.rate=cover.nb
county.id<-data.frame(county.id, cover.rate)

pc=cbind(cover.rate-mean(cover.rate), prob.nb-mean(prob.nb))
```

```
fit<-princomp(pc,cor=T)
pca1=fit$scores[,1]
```

```
rcontr<-rpart.control(minisplit=25, minbucket=4, cp=0.01)
cpf.split<-rpart(pca1~cover.rate+prob.nb, control=rcontr )
```

```
printcp(cpf.split)
x11()
plot(cpf.split$cptable[,"nsplit"],cpf.split$cptable[,"xerror"],
type="l", xlab="Tree Size", ylab="X relative error")
```

```
x11()
cr=cover.rate
pp=prob.nb
rcontr<-rpart.control(minisplit=25, minbucket=4, cp=0.04)
```

```
pcp=rpart(pca1~cr+pp, control=rcontr)
plot(pcp)
text(pcp)
```

```
#---- The plot of classification
x11()
plot(prob.nb, cover.rate, xlab="Prensence Probability (pp)", ylab="Cover Rate (cr)")
lines(rep(0.1695, 101),rep(0:100)/101, col="red")
lines(rep(0.3967, 101),rep(0:100)/101, col="red")
lines(rep(40:100)/100, rep(.08,61), col="red")
```

```
#---- Classified each county into different stages and plot
prob.cut=c(0.1695, 0.3967)
cover.cut=c(.0814)
```

```
for ( i in 1:num.county){
  if (prob.nb[i]==0) pstage[i]="None"
}
```

```
#---- The map of invasion stages
stages.plot()
```

```
length(which(pstage=="None"))
length(which(pstage=="Introduction"))
length(which(pstage=="Colonization"))
length(which(pstage=="Establishment"))
length(which(pstage=="Spread"))
```

```
#---- The proportions of plots
proportion.plot()
box()
```

60
APPENDIX B

CONSTRUCTING SPATIAL WEIGHT MATRIX

AND AN ECOLOGICAL EXAMPLE
This study constructed an spatial weight matrix using the local spatial statistics $G_i$. The simulation study is a replication of Getis and Aldstadt (2004). Three types (Random data set, 2-Cluster data set, and 6-Cluster data set of 30 by 30 raster data set are simulated and each type is simulated 25 times. The dataset from Forest Inventory and Analysis dataset (2005-2006) from U.S. Forest Service, Northern Research Station is used as an example of the application of the modified spatial weight matrix.

Created by Weiming Yu, 10/21/2011

--- Loading Packages------------------------------------------------------------
library(maps)
library(mapproj)
library(maptools)
library(sp)
library(spdep)
library(spatstat)
library(stats)
library(rpart)

--- Self-defined Functions------------------------------------------------------

This function generate a series with 2 clusters with radius 8 and centered (10, 10), (20, 20)

```
class2.gen<-function(grid=grid.n, radius=6, centroid=c(280, 590), num.row=nrow, num.col=ncol)
{
  N<-num.row*num.col
  std.norm<-rnorm(nrow*ncol, 0, 1)
  nb5 <- dnearneigh(grid.n, 0, radius)

  small.nb<-c(nb5[[centroid[1]]], centroid[1])
  large.nb<-c(nb5[[centroid[2]]], centroid[2])

  len1<-length(small.nb)
  len2<-length(large.nb)

  cstd<-std.norm[order(std.norm)]
  cstd.small<-cstd[1:len1]
  cstd.large<-cstd[(N-len2+1):N]
  cstd.middle<-cstd[(len1+1):(N-len2)]

  cluster2.std<-rep(10000, 1, N)
  cluster2.std[small.nb]<-cl.order(cstd.small, std.norm)
  cluster2.std[large.nb]<-cl.order(cstd.large, std.norm)
  cluster2.std[-c(small.nb, large.nb)]<-cl.order(cstd.middle, std.norm)

  return(cluster2.std)
}
```

This function generate a series with 6 cluster with radius 2, 4, 6 and centered (14, 27), (27,14), (8, 22), (22, 8), (10, 10), (20, 20)

```
position.get<-function(radius=c(2, 4, 6), centroid=c(417, 794, 232, 638, 280, 590))
{
  c1<-c(centroid[1], centroid[3], centroid[5])
  c2<-c(centroid[2], centroid[4], centroid[6])
  t1<-c(floor(runif(1)*2)+1, floor(runif(1)*2)+1, floor(runif(1)*2)+1)
  o1<-which(t1==2)
  m1<-c1[o1]
  c1[o1]<-c2[o1]
  c2[o1]<-m1
```

62
# Form two data.frame
  a1 <- data.frame(c1, radius)
a2 <- data.frame(c2, radius)
a1[order(rnorm(3, 0, 1)),] <- a1
a2[order(rnorm(3, 0, 1)),] <- a2
return(list(a1, a2))
}

cluster6.gen <- function(grid = grid.n, radius = c(2, 4, 6), centroid = c(417, 794, 232, 638, 280, 590), num.row = nrow, num.col = ncol){
  N <- num.row * num.col
  std.norm <- rnorm(nrow * ncol, 0, 1)
  cstd <- std.norm[order(std.norm)]
  high.position <- position[[1]]
  low.position <- position[[2]]
  nb1 <- dnearneigh(grid, 0, low.position[1, 2])
  nb2 <- dnearneigh(grid, 0, low.position[2, 2])
  nb3 <- dnearneigh(grid, 0, low.position[3, 2])
  nb4 <- dnearneigh(grid, 0, high.position[1, 2])
  nb5 <- dnearneigh(grid, 0, high.position[2, 2])
  nb6 <- dnearneigh(grid, 0, high.position[3, 2])
  clus.nb1 <- c(nb1[[low.position[1, 1]]], low.position[1, 1])
  clus.nb2 <- c(nb2[[low.position[2, 1]]], low.position[2, 1])
  clus.nb3 <- c(nb3[[low.position[3, 1]]], low.position[3, 1])
  clus.nb4 <- c(nb4[[high.position[1, 1]]], high.position[1, 1])
  clus.nb5 <- c(nb5[[high.position[2, 1]]], high.position[2, 1])
  clus.nb6 <- c(nb6[[high.position[3, 1]]], high.position[3, 1])

  len1 <- length(clus.nb1)
  len2 <- length(clus.nb2)
  len3 <- length(clus.nb3)
  len4 <- length(clus.nb4)
  len5 <- length(clus.nb5)
  len6 <- length(clus.nb6)
  cstd1 <- cstd[1:len1]
  cstd2 <- cstd[(len1+1):(len1+len2)]
  cstd3 <- cstd[(len1+len2+1):(len1+len2+len3)]
  cstd4 <- cstd[(N-len4-len5-len6+1):(N-len5-len6)]
  cstd5 <- cstd[(N-len5-len6+1):(N-len6)]
  cstd6 <- cstd[(N-len6+1):N]
  cstd7 <- cstd[(len1+len2+len3+1):(N-len4-len5-len6)]

  cluster6.std <- rep(10000, 1, N)
  cluster6.std[clus.nb1] <- cl.order(cstd1, std.norm)
  cluster6.std[clus.nb2] <- cl.order(cstd2, std.norm)
  cluster6.std[clus.nb3] <- cl.order(cstd3, std.norm)
  cluster6.std[clus.nb4] <- cl.order(cstd4, std.norm)
  cluster6.std[clus.nb5] <- cl.order(cstd5, std.norm)
  cluster6.std[clus.nb6] <- cl.order(cstd6, std.norm)
  combine <- c(clus.nb1, clus.nb2, clus.nb3, clus.nb4, clus.nb5, clus.nb6)
  cluster6.std[combine] <- cl.order(cstd7, std.norm)
  return(cluster6.std)
}

cl.order <- function(data1, data2){
  ind1 <- match(data1, data2)
  ind2 <- ind[order(ind1)]
  clf <- data2[ind2]
  return(clf)
}

63
locg.get<-function(data, grid=grid.n, style="B", rlist=radius.sq){
  # This function create a local G matrix with different radius
  L<-length(rlist)
  LG<-matrix(0, length(grid[,1]), L)
  for(i in 1:L){
    nb <- dnearneigh(grid, 0, rlist[i])
    local.g2 <- localG(data, nb2listw(include.self(nb), style=style))
    LG[, i]<-local.g2
  }
  return(LG)
}

neighbour.get<-function(dc, grid=grid.n, Nt=N){
  nb<-include.self(dnearneigh(grid, 0, 1))
  for(i in 1:Nt){
    nb[[i]]<-include.self(dnearneigh(grid, 0, dc[[i]]))
  }
  return(nb)
}

radius.get<-function(ls, grid=grid.n, rlist=radius.sq, rmax=40){
  # This function find the autocorrelation radius for each cell based on local stats
  # ls is the list of local statistics; rlist is the list of possible radius
  # rmax is the maximum search radius
  n=length(grid[,1])
  dc.loc<-rep(0, n)
  for(i in 1:n) {
    flag<-1
    k<-1
    sg<-sign(ls[[i]])
    while( flag==1 ) {
      if(k==1 & k<rmax){
        dnb <- dnearneigh(grid, 0, rlist[k])
      } else {
        dnb <- dnearneigh(grid, rlist[k-1], rlist[k])
      }
      subloc.g<-sign(ls[dnb[[1]]])
      len.pos<-length(which(subloc.g<sg | subloc.g>sg))
      if( len.pos==0){
        dc.loc[i]<-rlist[k]
        k<-k+1
      } else { flag<-0 }
    }
  }
  return(dc.loc)
}

wei.get<-function(nblist, dc, LMat=locg.mat, grid=grid.n, rlist=radius.sq){
  # This function constructs a spatial weight matrix based on the local spatial statistics
  nb.len<-length(nblist)
  wlen<card(nblist)
  glist <- vector(mode="list", length=nb.len)
  glist <- nblists(nblist, grid)
  for( i in 1:nb.len ){
    if( dc[i]==0 ) { glist [[i]]<-rep(0,wlen[i])}
    if( dc[i]>0 ) {
      # for j in 1:wlen[i]{
      if(i==nblist[[i]][[1]] & glist [[i]][1]<0){
        if(i==nblist[[i]][[1]] & glist [[i]][1]<0){
          if(i==nblist[[i]][[1]] & glist [[i]][1]<0){
            dij<-0
            index<-nblist[[i]][j]
            dji=sqrt( (grid[i, 1]-grid[index, 1])^2 +(grid[i, 2]-grid[index, 2])^2 )
            if( dji==dij1 ) {
              if( dji==dij1 ) {glist[[i]][j]<-1}
            } else {
              glist[[i]][j]<-dji
            }
          } else {
            glist[[i]][j]<-dij1
          }
        } else {
          glist[[i]][j]<-dij1
        }
      } else {
        glist[[i]][j]<-dij1
      }
    }
if (dc[i]>dn1 & dij<=dc[i]) {
  id.ij<-match(dij, rlist)
  id.dc<-match(dc[i], rlist)
  locg.dc<-LMat[i, id.dc]
  locg.ij<-LMat[i, id.ij]
  glist[[i]][j]<- abs( (locg.dc-locg.ij)/(locg.dc-LMat[i,1]))
}
}
}

nb2lw<-nb2listw(nblist, glist=glist)
return(nb2lw)

radius.get2<-function(data, LMat, grid=grid.n, rlist=radius.sq) {
  # This function find the autocorrelation radius for each cell based on
  # local stats? rlist is the list of possible radius; rmax is the maximum
  # search radius; LMat is the local G* statistics matrix
  Nnum=length(grid[,1])
  rmax<-length(glist)-1
  dc.loc<-rep(0, Nnum)
  for (i in 1:Nnum)
    L<-2
    if( LMat[i,1]>0 ){
      while(L>0){
        if(LMat[i, L-1]<=LMat[i, L] & L<=rmax) {
          dc.loc[i]<-rlist[L]
          L<-L+1
        } else {
          L<-0
        }
      }
    }
    if( LMat[i,1]<0 ){
      while(L>0){
        if(LMat[i, L-1]>=LMat[i, L] & L<=rmax) {
          dc.loc[i]<-rlist[L]
          L<-L+1
        } else {
          L<-0
        }
      }
    }
  return(dc.loc)
}

wei2.get<-function(nblist, dc, LMat=locg.mat, k=1, grid=grid.n, rlist=radius.sq) {
  # This function constructs a spatial weight matrix based on the local spatial statistics
  nb.len<-length(nblist)
  wlen<-card(nblist)
  glist <- vector(mode="list", length=nb.len)
  glist <- nblistw(nblist, grid)
  for (i in 1:nb.len)
    if( dc[i]==0 ) glist[[i]]<-rep(0,wlen[i])
    if( dc[i]>0 ) {
      for (j in 1:wlen[i])
        if( i==nblist[[i]][j]) glist[[i]][j]<-0
        if(i>nblist[[i]][j]) | i<nblist[[i]][j] {
          dij<0
          index<-nblist[[i]][j]
          dij=sqrt( (grid[i, 1]-grid[index, 1])^2 +(grid[i, 2]-grid[index, 2])^2 )
}
if( dc[i]==dn1 ) {
    if( dij==dc[i] ) {glist[[i]][j]<-1}
}

if( dc[i]==dn1 & dij==dc[i] ) {
    id_ij <- match(dij, rlist)
    id_dc <- match(dc[i], rlist)
    locg_dc <- LMat[i, id_dc]
    locg_ij <- LMat[i, id_ij]
    glist[[i]][j] <- abs((locg_dc/dc[i]^k-locg_ij/dij^k)/(locg_dc/dc[i]^k-LMat[i,1]))
}

}

if(!is.na(sum(glist[[i]]))) {
    if( sum( glist[[i]] ) > 0 ) {
        glist[[i]] <- glist[[i]]/sum(glist[[i]])
    }
}

nb2lw <- nb2listw(nblist, glist=glist)
return(nb2lw)

radius.get3 <- function(data=prob, coordinate=cord, style="B", rlist=dist.list){
    d.len <- length(data)
    dc <- rep(0, d.len)
    max.g2 <- rep(0, d.len)
    for( i in 1:d.len ){
        flag <- 0
        j <- 500
        while(flag==0){
            nb <- dnearneigh(coordinate, 0, rlist[j])
            local.g2 <- localG(data, nb2listw(include.self(nb), style=style))
            if(local.g2[i]>=max.g2[i]){
                max.g2[i] <- local.g2[i]
                dc[i] <- rlist[j]
                j <- j+3
            } else {
                flag <- 1
            }
        }
        dc.max <- data.frame(dc, max.g2)
    }
    return(dc.max)
}

wei3.get <- function(data=prob, nblist=isnb, Dc.LG=rc.radius, k=1, grid=cord, rlist=dist.list){
    # This function constructs a spatial weight matrix based on the local spatial statistics
    dc <- Dc.LG[,1]
    locg.dc <- Dc.LG[,2]
    nb.len <- length(nblist)
    wlen <- card(nblist)
    glist <- vector(mode="list", length=nb.len)
    glist <- nbdists(nblist, grid)
    for( i in 1:nb.len ){
        if( dc[i]==0 ) { glist[[i]] <- rep(0, wlen[i])}
        if( dc[i]>0 ) {
            for( j in 1:wlen[i]){ 
                if( i==nblist[[i]][j] ) {glist[[i]][j]<-rep(0,wlen[i])}
                if( i==nblist[[i]][j] & i==nblist[[i]][j] ) {
                    dij <- 0
                    index <- nblist[[i]][j]
                    dij <- sqrt( (grid[i,1]-grid[index,1])^2+(grid[i,2]-grid[index,2])^2 )
                } else {
                    for( j in nblist[[i]][j] ) {
                        if( j==i ) {glist[[i]][j]<-0}
                    }
                }
            }
        }
    }
}

66
if (dc[i] == dn1) {
    if (dij == dc[i]) {
        glist[[i]][j] <- 1
    }
    if (dc[i] > dn1 & dij <= dc[i]) {
        locg.ij <- localG(data, nb2listw(include.self(dnearneigh(grid, 0, dij)), style="B"))[i]
        glist[[i]][j] <- abs((locg.dc[i]/dc[i]^k - locg.ij/dij^k)/(locg.dc[i]/dc[i]^k-zero.g2[i]))
    }
}
}

if(!is.na(sum(glist[[i]]))) {
    if (sum(glist[[i]]) > 0) {
        glist[[i]] <- glist[[i]]/sum(glist[[i]])
    }
}

nb2lw <- nb2listw(nblist, glist=glist)
return(nb2lw)

stat.get <- function(data) {
    mean <- mean(data)
    max <- max(data)
    min <- min(data)
    SD <- sqrt(var(data))
    stat <- data.frame(mean, max, min, SD)
    return(stat)
}

TranMatrix <- function(data) {
    M <- matrix(0, 4, 6)
    value <- data[[2]]
    for (i in 1:3) {
        M[, 2*i-1] <- t(value[4*(i-1)+rep(1:4)])
        M[, 2*i] <- t(value[4*(i+2)+rep(1:4)])
    }
    return(M)
}

sw.simulation <- function(type="random", p=1, Nsim=nsimu) {
    nrow <- 30
    ncol <- 30
    AIC <- rep(0, Nsim)
    rho <- rep(0, Nsim)
    MI <- rep(0, Nsim)
    AIC2 <- rep(0, Nsim)
    rho2 <- rep(0, Nsim)
    MI2 <- rep(0, Nsim)

    for (L in 1:Nsim) {
        #---- generate data
        data <- rep(0, N)
        if(type == "random") data <- morm(N, 0, 1)
        if(type == "cluster2") data <- cluster2.gen()
        if(type == "cluster6") data <- cluster6.gen()

        #---- construct spatal weight matrix
        locg.mat <- locg.get(data) # local G* for different radius
        dc <- rep(0, 1, nrow*ncol)
        dc <- radius.get(data, locg.mat)

        #---- Define the dummy variables
        dummy <- rep(0, N)
        dummy[which(dc==0)] <- 1
    }
#---- generate the neighbour list
nb<-neighbour.get(dc)

#---- create the spatial weight matrix based on the local G*
weight<-wei.get(nblist=nb)
weight2<-wei2.get(nblist=nb, dc=dc, k=1)

#---- Spatial Lag Model
SLM<-lagsarlm(data~dummy, listw=weight, method="eigen", quiet=FALSE)
SLM2<-lagsarlm(data~dummy, listw=weight2, method="eigen", quiet=FALSE)

AIC[L]<-2*(SLM$parameters-SLM$LL)
rho[L]<-SLM$rho
Moran<--moran(SLM$lm.model$residuals, listw=weight, length(nb), Szero(weight))
MI[L]<--unlist(MoranSLM[I][1])
AIC2[L]<-2*(SLM2$parameters-SLM2$LL)
rho2[L]<-SLM2$rho
Moran<--moran(SLM2$lm.model$residuals, listw=weight2, length(nb), Szero(weight2))
MI2[L]<--unlist(MoranSLM[I][1])

AIC.stat<-stat.get(AIC)
rho.stat<-stat.get(rho)
MI.stat<-stat.get(MI)
AIC2.stat<-stat.get(AIC2)
rho2.stat<-stat.get(rho2)
MI2.stat<-stat.get(MI2)

d1<-data.frame(AIC, rho, MI, AIC2, rho2, MI2)
d2<-data.frame(AIC.stat, rho.stat, MI.stat, AIC2.stat, rho2.stat, MI2.stat)

return(list(d1, d2))

sim.plot<-function(data, color.type="gray"){  
br<-(ceiling(max(data))+1)
if(color.type=="gray"){  
  brks<--seq(0,2*br-1,0.1)
  cm.col<-hcl(seq(-225, 135, length=length(brks)-1))
  } else if(color.type=="color"){  
  brks<--seq(-5,5,1)
  cm.col<-hcl(seq(-45, 315, length=length(brks)-1))
  }
  x11()
  image.plot(x, y, t(matrix(data, nrow=n, ncol=n), byrow=TRUE), breaks=brks, col=cm.col, asp=1,xlab="x", ylab="y")
  box()
}

#--- Initial setting
n<30
nsimu<25
dn1<1
dc<4
nrow<n
ncol<n
N<nrow*ncol

#---- define the list of possible radius values
radius.sq<--rep(0, 100)
for( i in 0:12)
  for( j in 0:12)
    radius.sq[i*10+j] <<- i^2+j^2

radius.sq<--radius.sq[order(radius.sq)]
radius.sq <- radius.sq[radius.sq > 0]
radius.sq <- c(0, sqrt(radius.sq[!duplicated(radius.sq)]))
x <- rep(0:(n-1)) + 0.5
y <- rep(0:(n-1)) + 0.5

#---- Generate 30 X 30 grid ####
x.grid <- rep(0:(n-1), n) + 0.5
y.grid <- rep(seq(0,n-1), c(rep(n, n)))) + 0.5
grid.n <- cbind(x.grid, y.grid)

#---- Simulation Study

#---- Generate random cluster: 30 X 30
std.norm <- rnorm(N, 0, 1)
sim.plot(std.norm, "color")
locg.mat <- locg.get(std.norm, style="B") # local G* for different radius
dc <- rep(0, 1, nrow*ncol)
dc <- radius.get2(std.norm, locg.mat)
sim.plot(dc)

#---- Generate data with 2 clusters
cluster2.data <- cluster2.gen()
sim.plot(cluster2.data, "color")
locg.mat <- locg.get(cluster2.data, style="B") # local G* for different radius
dc.cluster2 <- rep(0, 1, nrow*ncol)
dc.cluster2 <- radius.get2(cluster2.data, locg.mat)
sim.plot(dc.cluster2)

#---- Generate data with 6 clusters
cluster6.data <- cluster6.gen()
sim.plot(cluster6.data, "color")
locg.mat <- locg.get(cluster6.data, style="B") # local G* for different radius
dc.cluster6 <- rep(0, 1, nrow*ncol)
dc.cluster6 <- radius.get2(cluster6.data, locg.mat)
sim.plot(dc.cluster6)

# Simulation Results

Nsims <- 25
AIC.c6 <- rep(0, Nsims)
rho.c6 <- rep(0, Nsims)
MI.c6 <- rep(0, Nsims)
AIC2.c6 <- rep(0, Nsims)
rho2.c6 <- rep(0, Nsims)
MI2.c6 <- rep(0, Nsims)

random.sim <- sw.simulation()
random.result <- TranMatrix(random.sim)

cluster2.sim <- sw.simulation(type="cluster2")
cluster2.result <- TranMatrix(cluster2.sim)
cluster2.result

cluster6.sim <- sw.simulation(type="cluster6")
cluster6.result <- TranMatrix(cluster6.sim)
cluster6.result

result <- data.frame(random.result, cluster2.result, cluster6.result)
result

69
# Invasive Species_Multiflora Rose Study

```r
ispath<="F:\Thesis\"
k<-1
isdata<-read.csv(paste(ispath, "InvasiveSp_MFR.csv", sep=""))
names(isdata)

centroid<-read.csv(paste(ispath, "centroid.csv", sep=""))
names(isdata)

xcord<-centroid$x
ycord<-centroid$y
cord<-cbind(xcord, ycord)
cord.len<-length(xcord)
prob<-isdata$prob.nb
dist.list<-rep(0, cord.len^2/2)
for(i in 1:cord.len){
  for(j in (i+1):cord.len){
    dist.list[k]<-sqrt((xcord[i]-xcord[j])^2+(ycord[i]-ycord[j])^2)
    k<-k+1
  }
}
dist.list<-dist.list[!duplicated(dist.list)]
dist.list<-dist.list[order(dist.list)]
dc.radius<-radius.get3()
cradius
dn1<-0.00001

nb<-include.self(dnearneigh(cord, 0, dist.list[1000]))
## rc.radius<-rc2.radius
# rc.radius<-dc.radius

zero.g2<-localG(prob, nb2listw(include.self(dnearneigh(cord, 0, 0)), style="B"))
isnb<-neighbour.get(rc.radius[,1], grid=cord, Nt=649)

#---- create the spatial weight matrix based on the local G*
weight2.is<-wei3.get(k=1)
weight.is<-wei3.get(k=0)

#---- Define the dummy variables
dummy<-matrix(0, 649)
dummy[which(rc.radius[,1]==0)]<-1

irpf<-isdata$interstate.density/(isdata$percent.forest+1)*1000
irpf<-irpf/max(irpf)
srpf<-isdata$stateroad.density/(isdata$percent.forest+1)*1000
srpf<-srpf/max(srpf)

range(irpf)
range(srpf)

#---- Spatial Lag Model
is3.SLM<-spautolm(prob.nb~Lon.x+Lat.y+ub+ub.sq+irpf+srpf, data=isdata, listw=weight2.is)
summary(is3.SLM, Nagelkerke=T)

is4.SLM<-spautolm(prob.nb~Lon.x+Lat.y+ub+ub.sq+irpf+srpf, data=isdata, listw=weight.is)
summary(is4.SLM, Nagelkerke=T)

#---- Spatial Lag Model
is3.SLM<-spautolm(prob.nb~Lon.x+Lat.y+ub+ub.sq+irpf, data=isdata, listw=weight2.is)
summary(is3.SLM, Nagelkerke=T)
isres.SLM<-is3.SLM$fit$residuals
moran.test(isres.SLM, weight2.is, zero.policy=TRUE)
```