Geographical Analysis ISSN 0016-7363

Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters

Jared Aldstadt,[1,2] Arthur Getis[1]

[1]Department of Geography, San Diego State University, San Diego, CA, [2]Department of Geography, University of California, Santa Barbara, CA

The creation of a spatial weights matrix by a procedure called AMOEBA, A Multidirectional Optimum Ecotope-Based Algorithm, is dependent on the use of a local spatial autocorrelation statistic. The result is (1) a vector that identifies those spatial units that are related and unrelated to contiguous spatial units and (2) a matrix of weights whose values are a function of the relationship of the i*th spatial unit with all other nearby spatial units for which there is a spatial association. In addition, the AMOEBA procedure aids in the demarcation of clusters, called ecotopes, of related spatial units. Experimentation reveals that AMOEBA is an effective tool for the identification of clusters. A comparison with a scan statistic procedure (SaTScan) gives evidence of the value of AMOEBA. Total fertility rates in enumeration districts in Amman, Jordan, are used to show a real-world example of the use of AMOEBA for the construction of a spatial weights matrix and for the identification of clusters. Again, comparisons reveal the effectiveness of the AMOEBA procedure.*

Introduction

In recent years, two problems have been especially troublesome for spatial analysts: (1) the difficulty in creating a defensible spatial weights matrix for use in spatial autoregressive (SAR) models (the context for the problem is outlined in Anselin 1988) and (2) finding spatial clusters on a map, that is, identifying ''geographically bounded group[s] of occurrences of sufficient size and concentration to be unlikely to have occurred by chance'' (Knox 1989). In an edited volume on this subject (Alexander and Boyle 1996), the many sides to this problem are discussed. In this article, we introduce a methodology that addresses these problems simultaneously and gives results that appear to be an improvement over current techniques used for their solutions.

Correspondence: Jared Aldstadt, Department of Geography, San Diego State University, San Diego, CA 92182
e-mail: aldstadt@rohan.sdsu.edu

Submitted: July 7, 2005. Revised version accepted: February 3, 2006.

Our method is in the form of an algorithm, *A Multidirectional Optimal Ecotope-Based Algorithm* (AMOEBA), which is a design for the construction of a spatial weights matrix using empirical data that can also simultaneously identify the geometric form of spatial clusters. It is *multidirectional* in that it searches for spatial association in all specified directions from one or more "seed" spatial units. It is *optimum* in the sense that the scale is the finest (most local) at revealing all of the spatial association that is subsumed in the data. We use the term *ecotope-based* to represent the technique's emphasis on finding subregions of spatial association within the entire spatial data set. In the environmental literature, a specialized region within a larger region is termed a habitat or ecotope. In effect, we introduce an *algorithm* for finding ecotopes that are spatial clusters of association among data points. Very often, these ecotopes are spatially irregular, or amoeba-like. The algorithm shares some common features with region-growing techniques from the image segmentation literature (Adams and Bischof 1994). The usual goal of these algorithms is to partition a large number of pixels exhaustively into homogeneous regions from a much smaller number of seed pixels. The emphasis here is on statistically significant clusters of high and low values and is not necessarily exhaustive.

AMOEBA is based on a principle developed earlier (Getis and Aldstadt 2004) that spatial structure can be considered in a two-part framework that separates spatially associated data from nonspatially associated data. Fundamental to AMOEBA is a single type of local statistic that is used to test the existence of a spatial association between nearby spatial units. In this article, for demonstration purposes, we use the local *G* statistic (G_i^*) of Getis and Ord (1992) and Ord and Getis (1995). The subscript *i* implies that the statistic's focus is on a particular spatial unit. We demonstrate AMEOBA's usefulness by a comparison with a procedure based on the scan statistic called SaTScan. The evaluation uses georeferenced total fertility rate data from Amman, Jordan.

The literature on the creation of spatial weights matrices (**W**) is extensive. Getis and Aldstadt (2004) identify over a dozen different types of **W**. The simplest are contiguity-type matrices, while among the most complex are those based on geostatistical models. In between these extremes are a host of different distance-related formulations. The point to be made, however, is that the need for a **W** in research is based on the notion that any spatial model must account for the spatial association extant within any region that has been divided into its constituent parts.

Those developing spatial models have as their viewpoint of **W** one of the following three types of representations:

1. A theoretical notion of spatial association, such as a distance decline function.
2. A geometric indicator of spatial nearness, such as a representation of contiguous spatial units.
3. Some descriptive expression of the spatial association within a set of data, such as an empirical variogram function.

For viewpoint 1, modelers argue that a spatial weight matrix is exogenous to any system and should be based on a preconceived matrix structure. A typical theoretical formulation for **W** is based on a strict distance decline function such as $1/d_{ij}^2$ where i and j are the centroids of spatial units. As little theory is available for the creation of these matrices, many researchers follow viewpoint 2, that is, they resort to geometric **W** specifications, such as a contiguity matrix (1's for contiguous neighbors and 0's for noncontiguous neighbors), reasoning that it is the nearest neighboring spatial units that bear most heavily on each other and thus represent the spatial association in a given set of georeferenced data. They argue that, with respect to an absence of theory, a simple near-neighbor approach appears to be reasonable (see, especially, Bartels 1979). For viewpoint 3, modelers 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 **W** matrix from the observed spatial associations. As a result, models based on this type of endogenous specification have limited explanatory power, the limit being the reference region. AMOEBA is based on viewpoint 3.

The rationale for AMOEBA is simply that, as viewpoints 1 and 2 may not represent the reality that is embodied in their study data, researchers are better advised to adopt viewpoint 3 by creating **W** from their already existing data. At the very least, they can argue that the complexities of spatial association within their data will be included in any SAR model. In a sense, we have differentiated strictly explanatory models of viewpoints 1 and 2 from the descriptive models of viewpoint 3. Note that the entire field of geostatistics, a field that emphasizes forecasting, is based on viewpoint 3 (Cressie 1993).

In the following section, we outline the way AMOEBA produces a **W** matrix. This leads to representations for spatially and nonspatially associated observations in addition to the **W** matrix. Next, we demonstrate the use of AMOEBA in a SAR model environment using data from Amman, Jordan. Then, by means of an example using artificial data, we show how AMOEBA can act as a spatial cluster identifier. We then compare our methodology with the spatial cluster identification technique SaTScan (Kulldorff 1997). Finally, closing comments are offered.

Creating W with AMOEBA

AMOEBA is an algorithm for creating a spatial weights matrix from univariate, areal spatial data. In our explanation of AMOEBA we use the Getis–Ord local statistic G_i^* (Ord and Getis 1995). For a given location i, the statistic G_i^* is defined as

$$G_i^* = \frac{\sum_{j=1}^{N} w_{ij}x_j - \bar{x}\sum_{j=1}^{N} w_{ij}}{S\sqrt{\frac{\left[N\sum_{j=1}^{N} w_{ij}^2 - \left(\sum_{j=1}^{N} w_{ij}\right)^2\right]}{N-1}}} \qquad (1)$$

where N is the number of spatial units, x_j is the value of the phenomenon of interest at location j, $\bar{x}$ is the mean of all the values, and

$$S = \sqrt{\frac{\sum_{j=1}^{N} x_j^2}{N} - (\bar{x})^2}$$

w_{ij} is an indicator function that is one if unit j is in the same designated region as unit i and zero otherwise.

The null hypothesis for a test based on this statistic is that there is no association between the value found at a site and its neighbors within the designated region. The expected value under the null hypothesis is 0, and the variance is approximately 1. Therefore, the G_i^* statistic is distributed as a standard normal variate. It should be noted that, with minor procedural modifications, other local statistics such as Anselin's local Moran's I_i (Anselin 1994) or the spatial scan statistic (Kulldorff 1997) can be used in the AMOEBA procedure. Point data, however, can be used only if the points are aggregated into areal spatial units.

At the outset of the AMOEBA procedure, we compute the G_i^* value for the spatial unit i itself. This value is denoted $G_i^*(0)$ and the ecotope consists of just the ith unit. A $G_i^*(0)$ value greater than zero indicates that the value at location i is larger than the mean of all units and, correspondingly, a value less than zero indicates that the value at location i is smaller than the mean.

The next step is to compute the $G_i^*(1)$ value for each region that contains i and all combinations of its contiguous neighbors (see Fig. 1). For an ecotope with four contiguous neighbors, $\binom{4}{4} + \binom{4}{3} + \binom{4}{2} + \binom{4}{1} = 15$ different regions are evaluated. If $G_i^*(0)$ is greater or less than zero, the combination that maximizes the statistic $G_i^*(1)$ absolutely becomes a new high- or low-value ecotope. At each

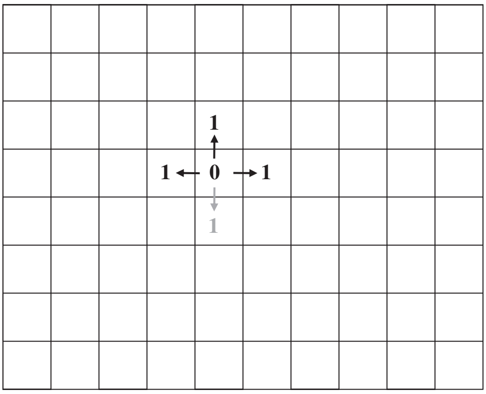

Figure 1. Stage one in AMOEBA determination: the bold arrows represent those links to units that are to be included in the ecotope. The light arrow indicates a link to a contiguous neighbor that will not be included in the ecotope.

succeeding step, contiguous units that are not included in the ecotope are eliminated from further consideration. Likewise, units included in the ecotope remain in the ecotope. Subsequent steps evaluate all combinations of contiguous neighbors and new members of the ecotope are identified. This process continues for k number of links, $k = 2, 3, 4, \ldots, \max$ (see Fig. 2). The final ecotope ($k_{\max}$) is identified when the addition of any set of contiguous units fails to increase the absolute value of the G_i^* statistic. Fig. 3 shows a complete AMOEBA ecotope in a raster setting. The maximum number of links in this case is five ($k_{\max} = 5$).

The results of the AMOEBA procedure are then used to construct **W** by using equation (2). All rows of **W** are row standardized, that is, each row is equated proportionally to sum to 1. This implies that the elements of **W** are relative within rows and not between rows. By convention, the diagonal elements (w_{ii}) are set to zero. The weight calculation for entry into the ith row of **W** is given by

$$
\begin{array}{ll}
\text{When } k_{\max} > 1 & \\
w_{ij} = \{P[z \le G_i^*(k_{\max})] - P[z \le G_i^*(k_j)]\}/ & \\
\qquad \{P[z \le G_i^*(k_{\max}) - P[z \le G_i^*(0)]\}, & \text{for all } j \text{ where } 0 < k_j \le k_{\max}, \\
w_{ij} = 0, & \text{otherwise;} \\
\text{when } k_{\max} = 1 & \\
w_{ij} = 1, & \text{for all } j \text{ where } k_j = 1, \\
w_{ij} = 0, & \text{otherwise;} \\
\text{when } k_{\max} = 0 & \\
w_{ij} = 0, & \text{for all } j.
\end{array}
\tag{2}
$$

Read k_j as the number of links connecting i and j in the ecotope. The probabilities in equation (2) are the cumulative probability associated with the G_i^* value. Thus, the numerator in equation (2) represents the area under the standard normal

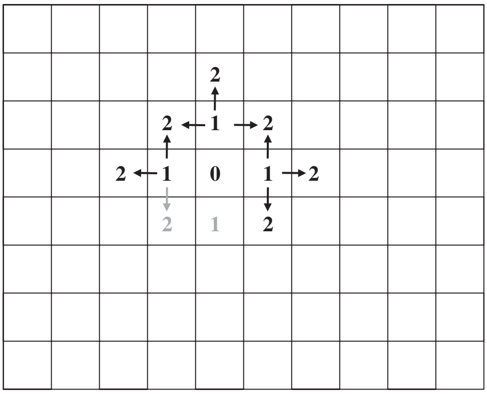

Figure 2. Stage two in AMOEBA determination: the bold numbers represent those cells that are to be included in the ecotope at stage 2.

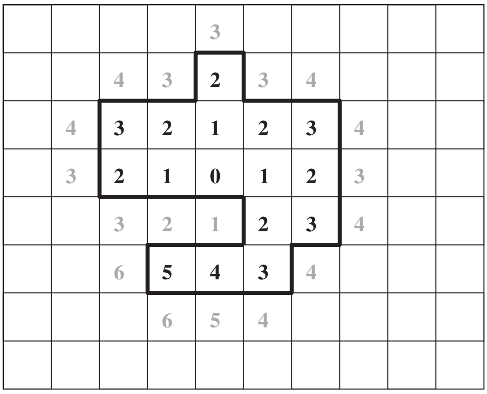

Figure 3. Completed AMOEBA pattern: the dark boundary outlines the final shape of the ecotope.

curve between $G_i^*(k_{\max})$ and $G_i^*(k_j)$. The values of w_{ij}, therefore, decrease as the number of links between units i and j increase.

When the ecotope contains only units one link away from the ith unit ($k_{\max} = 1$), each of these units is given a weight of 1. In this special case, there is no need for relative weighting of w_{ij} values.

When there is no spatial association between i and any j ($k_{\max} = 0$), the ith row of **W** is completely filled with zeroes. We differentiate between the "all zero" rows and those with any cell having a value greater than $w_{ij} = 0$. We create an $n \times 1$ vector of ones and zeroes where ones represent the zero rows and zeroes represent the remaining rows. Thus, the vector (**U**) acts as a dummy variable representing those observations having no spatial association with any other observation (see Fig. 4). We point out in Getis and Aldstadt (2004) that, under most circumstances, that is, when the number of zero rows in **W** is less than half of all rows, the nonsingularity requirement of the **W** matrix in a SAR equation is satisfied.

In sum, then, for each observation i, the set of j observations that maximizes the local statistic becomes a member of the ecotope together with the ith observation. The procedure ends when any combination of linked observations fails to increase the G_i^* value. We then enter all w_{ij} values into an $n \times n$ spatial weights matrix (**W**), where n represents the total number of observations. The distinguishing feature of this approach is its flexibility in identifying the spatial association of nearby units regardless of the configuration of those units.

We may generalize the AMOEBA approach in the following way[1]:

First partition $\boldsymbol{Y}$, the vector data values, into $\{\boldsymbol{Y}_c, \boldsymbol{Y}_0\}$ where subscript c denotes units that have an association with at least one other spatial unit, and 0 denotes spatial units with no such association. A spatial autoregressive model may be written as

$$\begin{bmatrix} \boldsymbol{Y}_c \\ \boldsymbol{Y}_0 \end{bmatrix} = \alpha \begin{bmatrix} 1_c \\ 1_0 \end{bmatrix} + \rho \begin{bmatrix} \mathbf{W}_{cc} & \mathbf{W}_{c0} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{Y}_c \\ \boldsymbol{Y}_0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_c \\ \varepsilon_0 \end{bmatrix} \tag{3}$$

$$\mathbf{U} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{W} = \begin{bmatrix}
0 & w_{1,2} & w_{1,3} & w_{1,4} & w_{1,5} & w_{1,6} & w_{1,7} & w_{1,8} & w_{1,9} & w_{1,10} & w_{1,11} & w_{1,12} & w_{1,13} & w_{1,14} \\
w_{2,1} & 0 & w_{2,3} & w_{2,4} & w_{2,5} & w_{2,6} & w_{2,7} & w_{2,8} & w_{2,9} & w_{2,10} & w_{2,11} & w_{2,12} & w_{2,13} & w_{2,14} \\
w_{3,1} & w_{3,2} & 0 & w_{3,4} & w_{3,5} & w_{3,6} & w_{3,7} & w_{3,8} & w_{3,9} & w_{3,10} & w_{3,11} & w_{3,12} & w_{3,13} & w_{3,14} \\
w_{4,1} & w_{4,2} & w_{4,3} & 0 & w_{4,5} & w_{4,6} & w_{4,7} & w_{4,8} & w_{4,9} & w_{4,10} & w_{4,11} & w_{4,12} & w_{4,13} & w_{4,14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_{7,1} & w_{7,2} & w_{7,3} & w_{7,4} & w_{7,5} & w_{7,6} & 0 & w_{7,8} & w_{7,9} & w_{7,10} & w_{7,11} & w_{7,12} & w_{7,13} & w_{7,14} \\
w_{8,1} & w_{8,2} & w_{8,3} & w_{8,4} & w_{8,5} & w_{8,6} & w_{8,7} & 0 & w_{8,9} & w_{8,10} & w_{8,11} & w_{8,12} & w_{8,13} & w_{8,14} \\
w_{9,1} & w_{9,2} & w_{9,3} & w_{9,4} & w_{9,5} & w_{9,6} & w_{9,7} & w_{9,8} & 0 & w_{9,10} & w_{9,11} & w_{9,12} & w_{9,13} & w_{9,14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_{11,1} & w_{11,2} & w_{11,3} & w_{11,4} & w_{11,5} & w_{11,6} & w_{11,7} & w_{11,8} & w_{11,9} & w_{11,10} & 0 & w_{11,12} & w_{10,13} & w_{11,14} \\
w_{12,1} & w_{12,2} & w_{12,3} & w_{12,4} & w_{12,5} & w_{12,6} & w_{12,7} & w_{12,8} & w_{12,9} & w_{12,10} & w_{12,11} & 0 & w_{12,13} & w_{12,14} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w_{14,1} & w_{14,2} & w_{14,3} & w_{14,4} & w_{14,5} & w_{14,6} & w_{14,7} & w_{14,8} & w_{14,9} & w_{14,10} & w_{14,11} & w_{14,12} & w_{14,13} & 0
\end{bmatrix}$$

Figure 4. Matrices **U** and **W**: note that when the *i*th row is 1 in **U**, all cells in the *i*th row are 0 in **W**. The AMOEBA procedure determines the w_{ij}.

where the various matrices and vectors are partitioned conformably. $\mathbf{W}_{cc}$ is the matrix of spatial association, and $\mathbf{W}_{c0}$ is the matrix of spatial association between units associated with at least one other unit and units having no spatial association with any other unit. The ρ is a regression coefficient that represents the strength of the spatial association in the study area taken as a whole. Each of the two ε vectors represents normally distributed residuals. This allows a different intercept ($\alpha+\beta$) for the "0" terms, against α for the "*c*" terms. Extensions to the model might include additional such terms to allow for more complex spatial structures.

If we write the likelihood for this model, we find that the estimator for ($\alpha+\beta$) is just $\sum \boldsymbol{Y}_0/n_0$ where n_0 is the number of "0" cells. The rest of the likelihood function reduces to the form for $\boldsymbol{Y}_c$ with $\boldsymbol{Y}_0$ as an "explanatory" variable. In Anselin's typology (1988), equation (3) may be considered a spatial lag autoregressive model.

AMOEBA, W, and total fertility levels in Amman, Jordan

As an example of the use of AMOEBA in its **W** context, we used a data set containing total fertility rates and many other social variables for enumeration districts ($n = 93$) in Amman, Jordan. See Weeks et al. (2004) for a complete description of this data set. Our goal is to:

1. estimate the parameters of a SAR model using an AMOEBA-generated spatial weights matrix and
2. compare the model in which AMOEBA is used with a misspecified ordinary least squares (OLS) model. The misspecification for OLS is due to the spatial autocorrelation contained within the data. In addition, we compare the AMOEBA SAR model with a model based on a row-standardized contiguity matrix for **W**.

Table 1 Parameter Estimates of Multiple Regression Models of Fertility Rates in Amman

	OLS		Contiguity		AMOEBA	
	β	*t*	β	*t*	β	*t*
Constant	4.42	6.27	4.54	6.50	1.74	3.55
Female education	− 0.14	− 14.34	− 0.14	− 13.05	− 0.11	− 11.55
Marital status	0.01	1.26	0.01	1.16	0.01	1.98
W coefficient (λ)			0.29	1.63	0.97	98.79
Nonspatial/spatial (α)					1.29	12.59
AIC	165.35		160.625		79.159	

AIC, Akaike information criterion; AMOEBA, A Multidirectional Optimum Ecotope-Based Algorithm; OLS, ordinary least squares.

For the AMOEBA SAR model, we use a spatial error formulation in which the **W** matrix is determined by the pattern of total fertility rates (TFR) values and row standardized. The model is of the form

$$\boldsymbol{Y} = \alpha\boldsymbol{U} + \boldsymbol{X}\beta + (\boldsymbol{I} - \lambda\mathbf{W})^{-1}\varepsilon \qquad (4)$$

where **U** is the vector representing the nonspatially associated data units (designated as 1) and the spatially associated units (indicated by 0). The ***X*** is the matrix of independent variables—percentage of females with higher education, percentage of females who are married—and ***Y*** is the dependent variable, TFR. The $(\boldsymbol{I} - \lambda\mathbf{W})^{-1}$ matrix represents that portion of the residuals explained by the spatial autocorrelation subsumed in **W**, and ε is the remaining vector of uncorrelated residuals. The parameters, α, β, and λ, to be estimated, indicate the strength of the variables. The software package SpaceStat (Anselin 1992) was used for this operation. Table 1 compares this model with the OLS and contiguity models. We use the Akaike information criterion (AIC) to aid in the evaluation of the three models.

The key findings are that the AMOEBA model: (1) identifies the importance of the independent variables and shows the extreme relevance of both the spatial (**W**) and the spatial/nonspatial (**U**) components of the model, (2) points out the weakness of both the OLS and contiguity formulations (AIC values are double that of the AMOEBA model—low AIC values represent good fitting models), and (3) indicates that the contiguity model fails to identify spatial autocorrelation as relevant ($t = 1.63$) to the model.

AMOEBA as a cluster identifier

For demonstration purposes, we create a spatial data map containing six obvious clusters of varying size: three clusters of high values and three clusters of low values (see Fig. 5). This illustration is designed to show how AMOEBA identifies arbitrarily shaped spatial clusters. The data set is made up of 900 observations in a 30 × 30 raster setup. The values constituting the observations were extracted from a normal

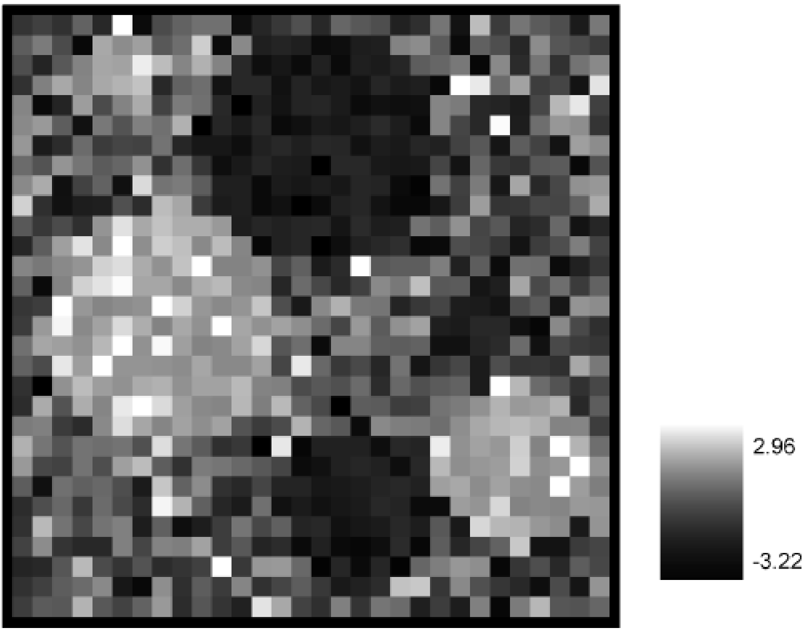

Figure 5. Artificial data set with six clusters: three with high values and three with low values.

distribution with mean 0 and variance equal to 1. Clusters are made up of values taken from the tails of the distribution. Those pixels not assigned to clusters were assigned values as random draws from the same normal distribution.

We present five examples of AMOEBA as a cluster identifier (see Fig. 6). Five pixels within the six-cluster data set are selected. These are as follows:

(a) the center of what clearly represents a cluster of high values;
(b) the inside edge of the same cluster;
(c) the edge of a small, irregularly shaped positive cluster within which is one pixel with a much lower value than the others;
(d) the center of a narrow, irregular cluster; and
(e) a low-value pixel surrounded by higher value pixels.

Note that Fig. 6a and b indicate that, within a cluster, no matter where the pixel location, center, or edge, the AMOEBA procedure delimits the same cluster. In addition, the cluster that is identified follows the visual outline of the constructed cluster. AMOEBA identifies the exact same pixels that we used to create the artificial cluster. The procedure does, however, include a few extra pixels (10 of them) not contained within the original artificial cluster of 113 pixels. This is due to the fact that, in the construction of the test map, a few pixels not earmarked for the artificial clusters, but having high values nonetheless, were randomly assigned to

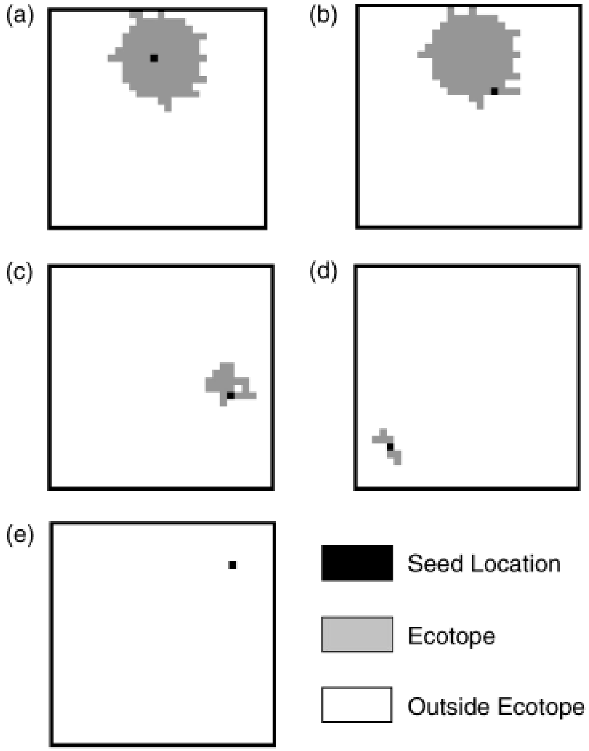

Figure 6. Five seed locations and ecotopes generated by AMOEBA. (a) A seed location near the center of a large cluster. (b) A seed close to the edge of a large cluster. (c) A seed within a small perforated cluster. (d) A seed that generates a linear ecotope. (e) A cell with a low value surrounded by high values is used as a seed.

be at the edge but outside of the artificial cluster. One cannot argue that these were mistakenly confused for the cluster as they are statistically significantly associated with nearby pixels.

In Fig. 6c, observe that AMOEBA, as expected, fails to create a cluster around the pixel with a purposely created low value. Fig. 6d emphasizes that, in their spatial complexity, the procedure easily identifies irregularly shaped clusters. Finally, in Fig. 6e, a pixel that does not conform to those around it (no spatial autocorrelation) shows up as an isolated pixel that is to be relegated to the **U** vector in the two-part spatial weights matrix.

In most cases, AMOEBA is able to start at any location within a cluster and go on to define the cluster, even if irregularly shaped. In the computer routine for finding clusters, we create an ecotope for each site. The ecotope with the highest $G_i^*(k_{\max})$ is selected as a cluster. Any ecotope overlapping with this "highest" cluster is removed from consideration. Then, the remaining ecotope with the next highest $G_i^*(k_{\max})$ value is selected. This process continues until no further ecotopes are identified.

The exact probability that each ecotope has arisen by chance is then evaluated using a Monte Carlo-type permutation test. A large number of random permutations of the data set are generated. These permutations involve randomly placing the N observed values among the N spatial units. For each of these permutations, the G_i^* statistic is calculated for the ecotope. The P value is then calculated as the rank of the observed data set divided by the number of Monte Carlo realizations plus one. Only those ecotopes with P values below some predesignated level of significance are considered as true clusters.

AMOEBA compared with SaTScan

One of the most popular and statistically sound cluster identifiers is based on the spatial scan statistic (Kulldorff 1997). A well-known software package, SaTScan, makes it easy to apply the procedure to both raster and vector data (Kulldorff et al. 1998; Kulldorff 2005). The following is a brief sketch of the nature of the spatial scan statistic and of SaTScan.

The spatial scan statistic is designed to find a spatial cluster that is unlikely to have occurred by chance. The likelihood can be based on any of a number of probability distributions. For aggregate spatial and spatiotemporal data, the Poisson distribution is used. The null hypothesis is that the count in each spatial unit is proportional to its population size. The likelihood ratio test statistic is

$$LR(R) = \left(\frac{c_R}{\mu_R}\right)^{c_R} \left(\frac{C - c_R}{C - \mu_R}\right)^{C - c_R} \tag{5}$$

where C is the total number of cases for the population, c_R is the number of cases in region R, and μ_R is the expected number of cases in region R. The most likely clusters are determined based on this maximum likelihood ratio statistic. An exact P value is calculated using a Monte Carlo procedure that involves generating a large number of realizations under the null hypothesis.

The SaTScan procedure uses a circular window (or kernel) on a map of the study region and allows the circle to move over it. The radius of the window continuously changes between zero and a specified upper limit. Each of the circles is considered to be a potential cluster. For each circle, one calculates the likelihood that one would find the observed sum of the values within the circle in consideration of the sum of values outside of the circle. Of all of the circles tried (a very large number), the one with the maximum likelihood is defined as the cluster that is the least likely to have occurred by chance.

In the following comparison with AMOEBA, we use an artificial data set more evocative of differences in cluster shape than that used in the earlier AMOEBA examples. We generated a 30 × 30 pixel image (900 pixels) with four clusters: two of high values and two of low values (see Fig. 7). These values were taken from the tails of a normal distribution (mean 100, standard deviation 25). The noncluster values were taken from a normal distribution with mean 100, but standard devi-

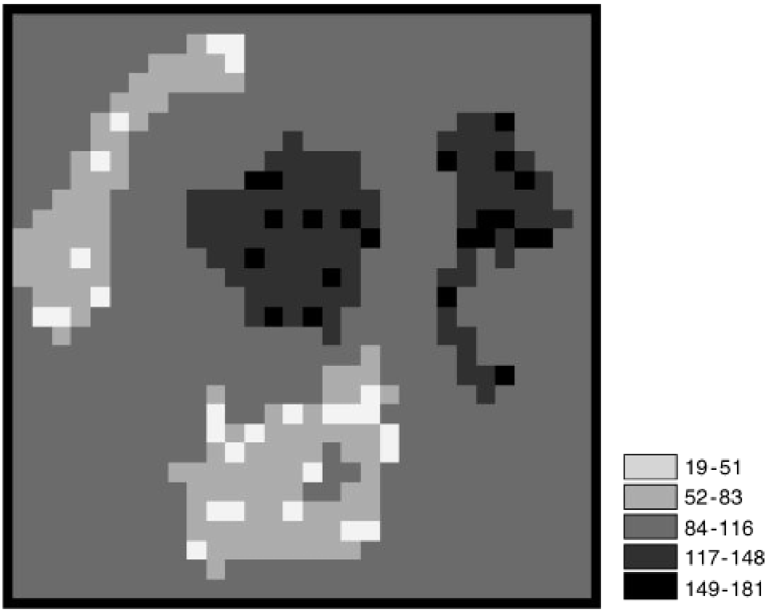

Figure 7. Four clusters with values for cells given in the legend.

ation 5 so that distinct clusters would obtain. We purposely selected ''difficult'' shapes for the clusters. One cluster shape is that of Norway (elongated). The second is that of Zimbabwe (compact). The next is Thailand (prorupt) and the fourth is South Africa (perforated by the country of Lesotho) (see Getis, Getis, and Fellmann 2006, p. 305, for further information on shape). We also purposely selected very obvious and distinct clusters. This is to ensure that the comparison is limited to the cluster search algorithms and not a comparison of the power of the two clustering statistics (G_i^* and spatial scan).

Utilizing the spatial scan statistic, the SaTScan software is designed to detect spatial or space-time clusters (a cluster of events that are grouped in space and time simultaneously) and to identify those that are statistically significant. ''A cluster detection test is able to both detect the location of clusters and evaluate their statistical significance without problems of multiple testing'' (Kulldorff 2005, p. 11). The options we chose for SaTScan were based on the nature of our data (cross-sectional areal data): purely spatial retrospective analysis, the Poisson probability model; the types of clusters sought (areas of high and low values); the Poisson model need for a base population (each pixel was set to 1000); and importance of clusters sought (secondary clusters were not reported if they overlapped a more statistically significant cluster). Fig. 8 shows the results of the SaTScan trial, and Fig. 9 identifies the misclassified pixels. Of the 900 pixels, 180 were misclassified. Of the 255 pixels in the artificially generated image that were contained in clusters, SaTScan properly identified 218, or 85%, but classified 143 of the 645 noncluster pixels as part of the clusters (22%). Overall, SaTScan's accuracy was 80%. The

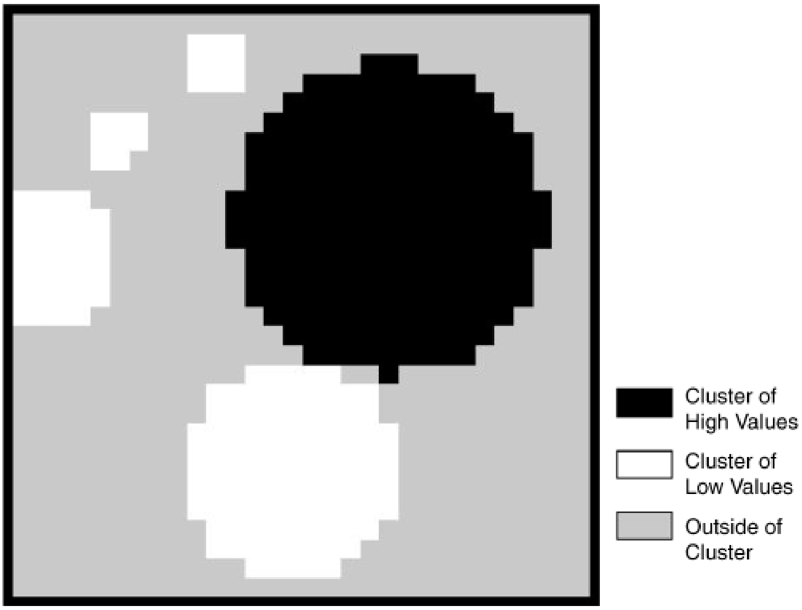

Figure 8. SaTScan clusters: one high-value and four low-value clusters. Compare with Fig. 7.

most likely reason for the high number of missed pixels is the circular shape of the SaTScan procedure. That is, it is basically constructed to find only compact clusters. Thus, in the case of the Zimbabwe shape, only four pixels in the cluster were misclassified.

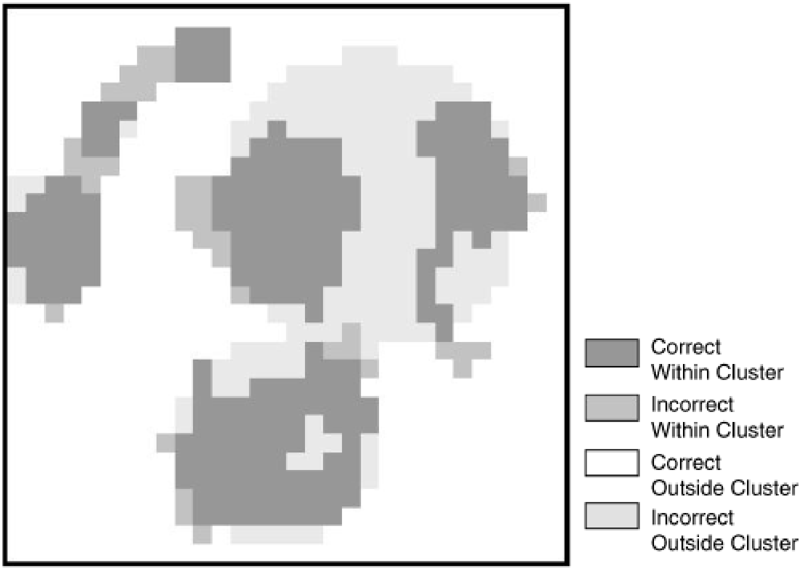

Figure 9. SaTScan: correctly and incorrectly classified pixels.

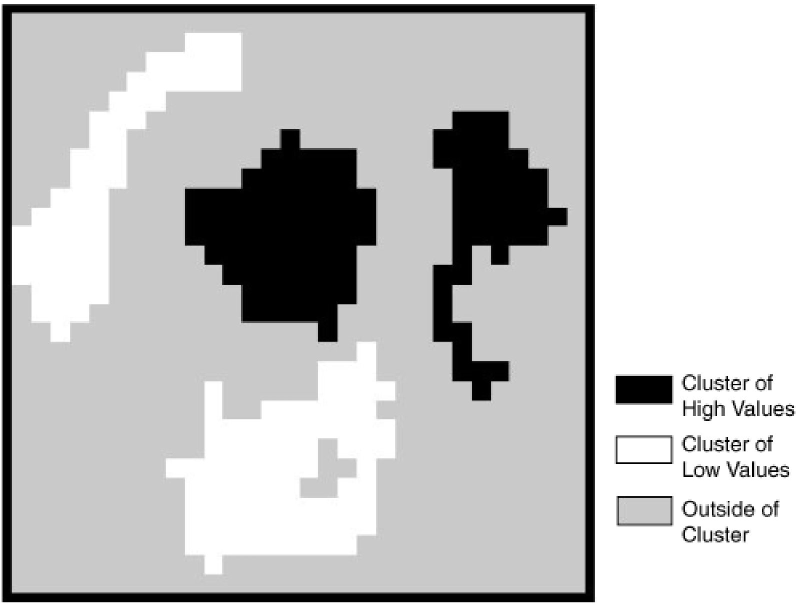

Figure 10. AMOEBA clusters: note the perfect correspondence with clusters shown in Fig. 7.

Further experimentation with SaTScan produced larger errors. For example, by choosing the option, ''no cluster centers in other clusters,'' the number of correct predictions was 66%, and when we chose ''no cluster centers in less likely clusters,'' the percentage correct was 70%. Using these less restrictive criteria resulted in more correct classifications of the pixels within clusters, but resulted in many more misclassifications of pixels outside of the clusters.

In contrast, Fig. 10 shows the results for AMOEBA using the G_i^* local statistic. There were no misclassified pixels; that is, AMOEBA was able to recreate the artificial clusters exactly.

For our second example, we show the TFR of Amman (Fig. 11a), the results of the AMOEBA cluster routine (Fig. 11b), and a comparison with the SaTScan clusters (''no geographical overlap'' option used, Fig. 11c). AMOEBA identifies an irregular cluster of high TFR rates in the east and southeast part of the city containing 33 EDs. In addition, there are two small clusters of high values. The low cluster is extremely irregular starting from the northwest and stretching into the center of the city (the well-to-do areas of the city). Note that 16 enumeration districts are not members of a cluster. In contrast, SaTScan again, because of its ''circular'' nature, finds compact clusters that do not always correspond to those identified by AMOEBA. Of the 33 high-value EDs found as the main cluster by AMOEBA, SaTScan missed 12 EDs and added two EDs that were not included in AMOEBA's determination.

In order to better understand any variability in clusters identified by AMOEBA, we considered seeds other than the one required, that is, ecotopes not with the highest $G_i^*(k_{\max})$ value. For the largest cluster in Amman, 25 of the 33 EDs as seeds

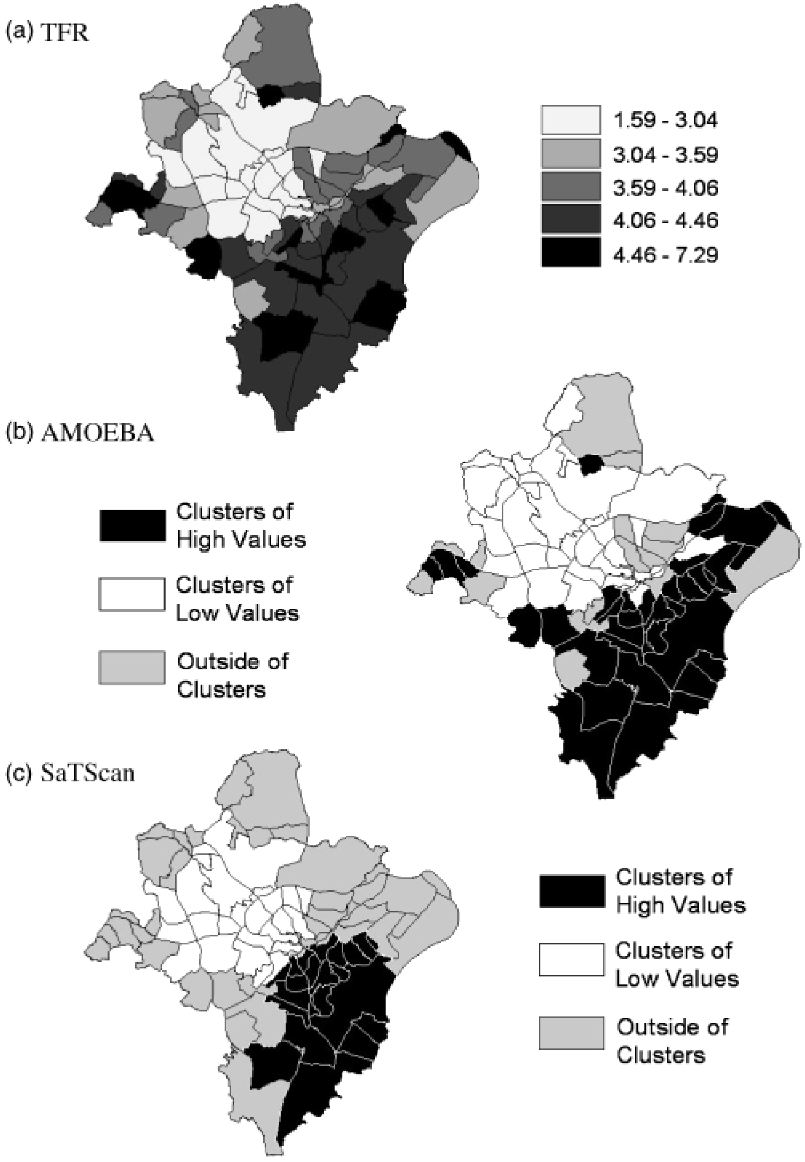

Figure 11. Census districts of Amman, Jordan. (a) Total fertility rates in 1994. (b) Clusters identified by AMOEBA. (c) Clusters identified by SaTScan.

produced the cluster seen in Fig. 11b with small variations (one produced a one spatial unit larger cluster and 24 produced a three-unit smaller cluster). The remaining eight seeds produced clusters of size one or two, a result that is obtained when extreme values result in high G_i^* values that are considerably higher than their near neighbors.

Summary and conclusions

As a device for determining the values in a spatial weights matrix, we have shown convincingly, although not unequivocally, that AMOEBA is a good tool for recognizing the manifestation of spatial association among nearby units. Significant improvement in model fit can be achieved by incorporating a weights matrix that is representative of the observed spatial association in the data. Using a vector that differentiates nonspatial from spatially associated units is a decided advantage when coming to grips with the complexity of interaction and noninteraction among spatial units.

In addition, AMOEBA is an effective device for finding clusters of weighted spatial units, even when the clusters are irregular in shape. Moreover, one need not use the same spatial unit to identify a cluster, but one must be prepared to see relatively small variations in the cluster size and shape depending on the seed location (this holds true under most circumstances).

In general, the AMOEBA and SaTScan routines differ mainly by their search algorithm. The circular nature of SaTScan tends to be inclusive of low-valued units in clusters of high values more so than AMOEBA. The rule that AMOEBA employs that possible overlapping clusters yield to the highest valued cluster implies that AMOEBA is the less inclusive. The rationale for this rule is that a statistically unimportant location would have to be included in a secondary overlapping cluster. Further, as AMOEBA's approach is a systematic step-by-step sequence in any direction, any low-valued location is quickly identified and eliminated from a possible cluster. In SaTScan's approach, it is the group of data sites within the circle that are evaluated together. This tends to generalize the cluster to include more low sites than in AMOEBA.

When compared with the SaTScan clustering routine, AMOEBA performs well. Apparently, algorithms that depend on particular shapes, such as circles or ellipses, are less successful at ferreting out the nuances of cluster configuration. In any case, it should be clear that SaTScan has much in common with AMOEBA; chief among the commonalities is the search for highly significant measures of spatial association.

Acknowledgements

We would like to express our appreciation for the funding that we received from the National Institutes of Health (AI034533) and the National Science Foundation (BCS-0095641 and BCS-0502020). We especially thank Keith Ord, Thomas Scott, John Weeks, and the anonymous referees for their support and comments.

Note

[1] This formulation was suggested to us by J. Keith Ord, Georgetown University.

References

Adams, R., and L. Bischof. (1994). ''Seeded Region Growing.'' *IEEE Transactions in Pattern Analysis and Machine Intelligence* 16(6), 641–47.

Alexander, F. E., and P. Boyle. (1996). *Methods for Investigating Local Clustering of Disease*. Lyon, France: IARC Scientific Publication.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht, Germany: Kluwer Academic Publishers.

Anselin, L. (1992). *SpaceStat, a Software Program for the Analysis of Spatial Data*. Santa Barbara, CA: National Center for Geographic Information and Analysis, University of California.

Anselin, L. (1994). ''Local Indicators of Spatial Association—LISA.'' *Geographical Analysis* 27, 93–115.

Bartels, C. P. A. (1979). ''Operational Statistical Methods for Analysing Spatial Data.'' In *Exploratory and Explanatory Statistical Analysis for Spatial Data*, edited by C. P. A. Bartels and R. H. Ketellapper. Boston: Martinus Nijhoff.

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

Getis, A., and J. Aldstadt. (2004). ''Constructing the Spatial Weights Matrix Using a Local Statistic.'' *Geographical Analysis* 36(2), 90–104.

Getis, A., J. Getis, and J. D. Fellmann. (2006). *Introduction to Geography*, 10th ed. Dubuque, IA: McGraw-Hill.

Getis, A., and J. K. Ord. (1992). ''The Analysis of Spatial Association by Distance Statistics.'' *Geographical Analysis* 24, 189–206.

Knox, G. (1989). ''Detection of Clusters.'' In *Methodology of Enquiries into Disease Clustering*, 17–22, edited by P. Elliott. London: Small Area Health Statistics Unit.

Kulldorff, M. (1997). ''A Spatial Scan Statistic.'' *Communications in Statistics: Theory and Methods* 26, 1487–96.

Kulldorff, M. (2005). *SaTScan User Guide* (version 5.1.1). http://www.satscan.org/

Kulldorff, M., K. Rand, G. Gherman, and G. Williams. (1998). *SaTScan v 2.1: Software for the Spatial and Space–Time Scan Statistics*. Bethesda, MD: National Cancer Institute.

Ord, J. K., and A. Getis. (1995). ''Local Spatial Autocorrelation Statistics: Distributional Issues and an Application.'' *Geographical Analysis* 27, 286–306.

Weeks, J. R., A. Getis, A. G. Hill, M. S. Gadalla, and T. Rashed. (2004). ''The Fertility Transition in Egypt: Intraurban Patterns in Cairo.'' *Annals of the Association of American Geographers* 94(1), 74–93.