Modelling Spatial Relations by Generalized Proximity Matrices

ANA PAULA DUTRA DE AGUIAR[1], GILBERTO CÂMARA[1] , ANTÔNIO MIGUEL VIEIRA MONTEIRO[1], RICARDO CARTAXO MODESTO DE SOUZA[1]

[1] Image Processing Division, National Institute for Space Research

Av. dos Astronautas, 1758 - 12227-001 - São José dos Campos , SP, Brazil

Abstract. One of the main challenges for the development of spatial information theory is the formalization of the concepts of *space* and *spatial relations*. Currently, most spatial data structures and spatial analytical methods used in GIS embody the notion of space as a set of *absolute locations* in a Cartesian coordinate system, thus failing to incorporate spatial relations, which are dependent on topological connections and fluxes between physical or virtual networks. To answer this challenge, we introduce the idea of a *generalized proximity matrix (GPM)*, an extension of the spatial weights matrix where the weights are computed taking into account both *absolute space* relations such as Euclidean distance or adjacency and *relative space* relations such as network connection. Using the GPM, two geographic objects (e.g. municipalities) are "near" each other if they are connected through a transportation or telecommunication network, even if thousands of kilometers apart or, using even more abstract concepts, if they are part of the same productive chain in a given economical activity. The generalized proximity matrix allows the extension of spatial analysis formalisms and techniques such as spatial autocorrelation indicators and spatial regression models to incorporate relations on relative space, providing a new way for exploring complex spatial patterns and non-local relationships in spatial statistics. The GPM can also be used as a support for map algebra operations and cellular automata models.

Keywords: Spatial relations, generalized proximity matrices, spatial analysis.

1 Introduction

The establishment of spatial information science as a scientific discipline requires it to possess a unique set of concepts, which are distinct from those used by other branches of science. In this respect, no concept is as crucial as the notion of *space* itself. After decades of research, the mathematical expression of *space* remains one of the most challenging problems in spatial information theory. Most spatial data structures used in GIS such as polygons and cells embody the notion of space as a set of *absolute locations* in a Cartesian coordinate system. Representation of *relative space* (the relation of a spatial object to other objects) in a GIS uses arc-node data structures. However, the models associated to arc-node data structures are usually completely unrelated to the analysis methods that use representations of absolute space, a separation that leads to a limited conception of space in geographical information system. This situation has led to much criticism both from within the GIS community and from non-practitioners ([1] [2] ([3]). Such critiques consider that flows of resources, information, organizational interaction and people are essential components of space. Therefore, efficient representation of such flows in connection with representation of absolute space is essential to achieve a realistic perspective of spatial relations [4]. Castells [5] views geographical space as a combination of “spaces of fixed locations and spaces of fluxes”, where the concept of ‘spaces of fixed locations’ represents spatial arrangements based on *absolute space*, and the concept of ‘spaces of fluxes’ indicates spatial arrangements based on *relative space*.

To achieve realistic computational models of spatio-temporal patterns, we need major advances in the representation of spatial relations. To take a motivational example, consider the process of modeling of land use change in the Brazilian Amazonia. Human occupation on the region drives this process, and has increased significantly in the last two decades. Models that project patterns of land use change in Amazonia have to consider that transportation networks (rivers and roads) play a decisive role in governing human settlement patterns. Figure 1 depicts urban settlements in Amazonia as white areas, and the road network in red lines. A realistic model for land use changes in the region has to take into account that the roads establish preferential directions for human occupation and land use changes. These relations would be impossible to capture in the isotropic neighborhoods prevalent in most spatial modeling techniques. Therefore, any spatial model that aims at understanding the processes in an area such as Amazonia requires flexible definitions of proximity that are able to capture action-at-a-distance.

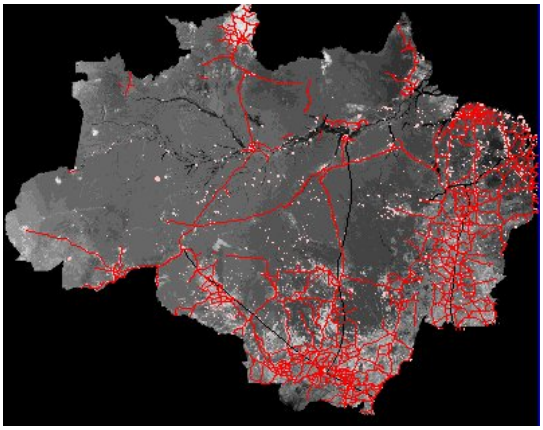

Figure 1 – Human settlements (white dots) and roads (red) in Amazonia

This paper proposes a model for expression of spatial relations, using a *generalized proximity matrix* (GPM). The GPM is an extension of the spatial weights matrix used in many spatial analysis methods[6], where the spatial relations are computed taking into account both *absolute space* relations such as Euclidean distance or adjacency and *relative space* relations such as topological connection on a network. This combination of *absolute space* and *relative space* has not been explored extensively before in the GIScience literature.

In this paper, we propose ways to combine the two notions of space and we illustrate the ideas with an example using data from colonization in the Brazilian Amazonia. The paper is organized as follows. Section 2 presents the basic definitions used in the paper. Section 3 describes the process of computation of a GPM. Section 4 demonstrates the use of the GPM in a case study of spatial relations in Amazonia. In section 5, we discuss the relation of the GPM to previous work.

2 Basic Definitions

Consider a set O of spatial objects whose geometrical representations are defined over a connected subset $S \subset \Re^2$. Examples of these objects include: (a) area regions whose boundaries are closed polygons; (b) cellular automata organized as sets of cells, whose boundaries are the edges of each cell; (c) point locations in two-dimensional space. A very large number of spatial analysis and spatial statistics functions over the set of objects O depend on the definition of a neighborhood. Therefore, given two objects o_i and o_j belonging to O, a basic question is "are these objects neighbors?"

The usual answer to this question is to consider that geographic objects are "close" to each other depending on their position in the *absolute space*, usually measured in terms of topological adjacency or Euclidean distance[7]. We denote the *neighborhood relation* between o_i and o_j by w_{ij}. Usual options for defining w_{ij} include:

- $w_{ij} = 1$, if o_i is topologically adjacent to o_j; $w_{ij} = 0$ otherwise. (1)
- $w_{ij} = 1$, if $\mathtt{distance}(o_i, o_j) < \delta$; $w_{ij} = 0$ otherwise. (2)

The set of all relations w_{ij} defines a *spatial weights matrix* W that represents the neighborhood relationships between objects in the set O. A large number of spatial analysis techniques use the spatial weights matrix, including moving averages, spatial autocorrelation indices and spatial regression methods [7]. To take one example, given one attribute of a set of spatial objects, we compute its local average by a weighted mean, where the weights are the relations w_{ij}:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} z_j}{\sum_{j=1}^{n} w_{ij}} \qquad (3)$$

In the above equation, $\hat{\mu}_i$ is the local mean, w_{ij} is the element of the weights matrix that relates objects o_i and o_j, and z_j is the value of the attribute for the object o_j.

We propose an extension of the usual definition of the spatial relation w_{ij} to include a combination of neighborhood measures in the *absolute* space and in the *relative (network)* space. We call the resulting matrix a *generalized proximity matrix* (GPM). In order to compute the GPM, we need *additional* information on the network relations between the objects in O. A graph G over the same connected subset S provides the connectivity information. The graph G is composed of a set of nodes N and a set of arcs A. Our definition of G includes different types of networks, including physical links (roads and rivers) and logical links (airline routes). This model of proximal space requires different representations of absolute and relative space using the sets S and G; we consider that configuration of the relative space cannot be derived satisfactorily from the boundaries of spatial objects alone.

There are many applications for the GPM. The most important ones involve using the GPM in spatial analysis and spatial modeling. The GPM inherits well-established formalisms and techniques such as spatial autocorrelation indicators and spatial regression models, which have been defined using spatial weights matrices.

3 Building Generalized Proximity Matrices

The computation of each element w_{ij} of the GPM requires two proximity measures: one associated to absolute space relations and a second one associated to relative space relations. We denote these functions by `proxabs` and `proxrel`, respectively. From the different possibilities for defining these functions, we will consider two alternatives: indicator functions and distance-based measures.

Indicator functions are functions take only values one (1) or zero (0), depending whether the chosen criteria is satisfied or not. In the case of `proxabs`, equations (1) and (2) are examples of indicator functions. The spatial weight is the logical union of both measures:

$$w_{ij} = \text{proxabs}(o_i, o_j) \cup \text{proxrel}(o_i, o_j) \quad (4)$$

As an alternative to indicator functions, *distance-based functions* use measure based on locations on Cartesian coordinates combined in a linear fashion:

$$w_{ij} = \alpha * \text{proxabs}(o_i, o_j) + \beta * \text{proxrel}(o_i, o_j) \quad (5)$$

Prior to the definition of the `proxrel` functions, we must first distinguish between two types of networks: (a) Networks in which the entrances and exits are restricted to its nodes, i.e., objects connect only at network nodes. Examples are railroads, telecommunication networks, banking networks, and productive chains. We denote these by *closed networks*; (b) Networks in which any location is entrance or exit point, i.e., objects connect at any node or arc coordinate. Examples are transportation networks such as roads and rivers. We denote these by *open networks*. For open networks, it is necessary to make use of the actual line coordinates that correspond to each arc in order to be able to compute the closest entrance/exit points from any arbitrary position.

The construction of the GPMs depends on the type of the network (open or closed) as explained in the next sections.

3.1 Proximity Measures in Open Networks

In the case of open networks, a spatial object connects to the network at any point of any of the arc, and not only at the nodes. The proposed criterion for GPM construction considers two basic parameters: (a) the maximum distance from an object to the network; (b) the maximum value of the shortest path between the two connection points of two spatial objects to the network.

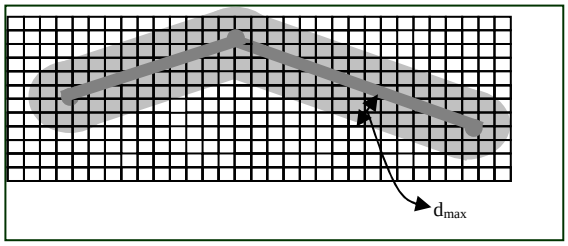

Figure 2. Schematic example of an open network.

Figure 2 presents a schematic example of an open network, where the spatial objects are regular cells and the maximum distance for connection to the network is dmax. We consider that two objects are neighbors in relative space when their centroids are inside the shaded area. Given a set `O` of spatial objects and a graph `G = (N, A)`, we use the following functions

- `cent::` $O \Rightarrow \Re^2$: denotes the function that calculates the centroid of a spatial object o_i;
- `dist::` $\Re^2 \times \Re^2 \Rightarrow \Re$: the function that calculates the distance of two points *x* and *y*;
- `shpath::` $\Re^2 \times \Re^2 \times G \Rightarrow \Re$: be the function that calculates the shortest path between two locations on a graph.
- `clspt::` $\Re^2 \times G \Rightarrow \Re^2$: be the function, that given a location in $\Re^2$ and a graph `G`, determines the closest point in the graph.

We can compute `proxrel`, as illustrated in Figure 3. Given a pair of objects o_i and o_j in `O` and a graph `G`, the closest locations in `G` to o_i and o_j are computed (p_i and p_j), and the shortest path between them is calculated. The objects are neighbors if: (a) the distance from their centroids to the network is smaller than a specified threshold (d_{max}); (b) the shortest path between p_i and p_j in the graph is smaller than a specified value (p_{max}). Two types of measurements are possible:

- Indicator function

```
bool proxrel (oi,oj, G) {
   pi = clspt (cent (oi), G);
   pj = clspt (cent (oj), G);
  if ((dist (cent(oi),pi)< dmax) AND
      (dist (cent(oj),pj)< dmax) AND
      shpath(pi, pj) < pmax)
      return TRUE;
   else return FALSE; }
```

- Distance-based function:

```
float proxrel (oi,oj, G) {
```

```
  p_i = clspt (cent (o_i), G);
  p_j = clspt (cent (o_j), G);
  return   (   1/shpath  (p_i,p_j,G)  +
            1/dist(cent(o_i),p_i)     +
            1/dist(cent(o_j),p_j));
}
```

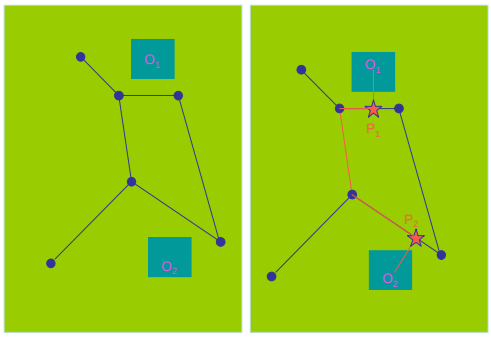

Figure 3. Schematic representation of algorithm for proximity measures in open networks.

A number of similar algorithms could be used to capture spatial relations on an open network, depending on factors such as the spatial configuration, the average length and the attributes of each arc. This added complexity might not be relevant in all applications, so it was not incorporated in the current discussion.

3.2 Proximity Measures in Closed Networks

In closed networks, the only entrance points are the nodes. In a similar way as in the case of open networks, two parameters are considered to identify connectivity between two spatial objects: the maximum distance from each object to the closest node and a maximum limit to the shortest path between the two nodes chosen as connection points, as illustrated in Figure 4. The cells whose centroid is inside the shaded area (green shade) are neighbors by the network criteria. Given these parameters, we can calculate the `proxrel` function in a similar way as in the previous section.

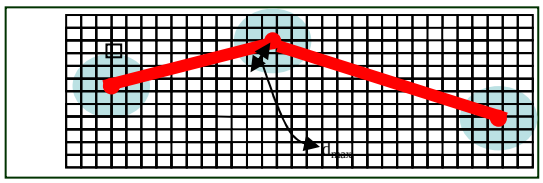

Figure 4. Schematic example of a closed network.

4 A Case Study on Spatial Analysis using the GPM

In this section, we exemplify the construction and usage of GPM in a case study. Our study area, shown in Figure 5, has approximately 260.000 km^2 and is located in the Brazilian Amazon rainforest, in the state of Pará. The study area was divided into regular cells of 625 km^2 (25 km per 25 km), and the attribute under analysis is the deforested area in each of these cells. We compared the local spatial autocorrelation indices obtained using GPMs constructed based on two different criteria: (a) proximity in absolute space (local adjacency) and (b) proximity in relative space (open network connections).

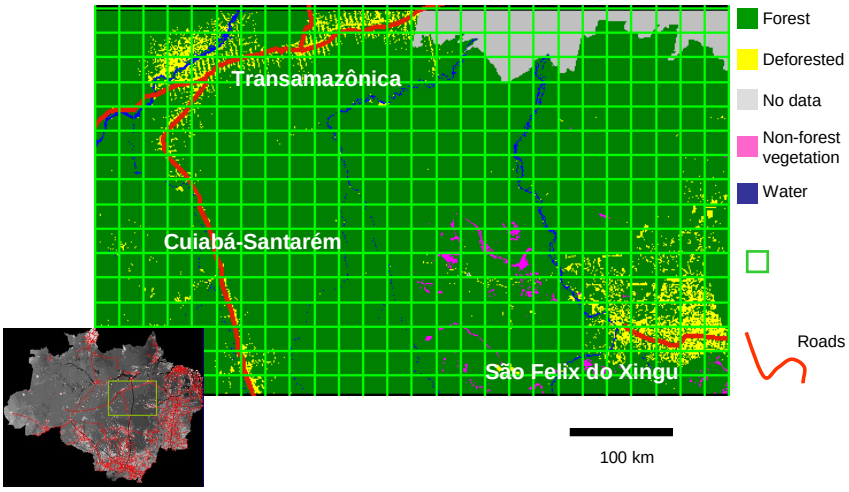

Figure 5. Study Area in Brazilian Amazonia, Pará State.

Two non-paved main roads, Transamazônica and Cuiabá-Santarém, cross the study area. The human occupation in the *Transamazônica* area dates mostly from the seventies; one can notice the "espinha de peixe" (fish spine) spatial pattern, caused by the lotting schema adopted by state planners in that area. The Cuiabá-Santarém region is a new frontier area, where the forest has been less disturbed, which has received a large recent influx of new settlers coming mainly from the south. In the southeast of the study area, there is a more consolidated agricultural region named São Felix do Xingu, which is also served by a non-paved road. The huge undisturbed forest area in the middle of the study area contains several conservation units and indigenous areas, but it is also being threatened by new settlers coming from the São Felix do Xingu region.

We are interested in studying the effects of road networks in the process of deforestation, taking into consideration that these effects are not homogeneous in space and time, given the differences in the territorial dynamics of agriculturally consolidated areas versus new frontier areas. We have used a local index of spatial autocorrelation (the Local Moran index) as an indication of the differences between the spatial patterns. The local analysis presented in this paper is an initial attempt in this direction. We calculate the Local Moran index for each object o_i, based on the product of one of its attributes (z_i) and the same attribute z_j of its neighbors [8]:

$$I_i = \frac{\sum_{j=1}^{n} w_{ij} z_i z_j}{\frac{\sum_{j=1}^{n} z_j^2}{n}} \qquad (10)$$

In the above formula, the GPM provides the weights w_{ij} and n is the number of neighbors. The closer the values of an object's attribute are those of its neighbors, the higher the index. Values around zero mean no correlation; higher positive values mean stronger positive correlation, and lower negative values mean stronger negative correlation. We also computed the statistical significance of the Local Moran Index, using 99 random permutations of the attribute values. Our goal was to analyze the behavior of such index given alternative neighborhood structures. We expected an increase in the indices when using the open network criterion (emphasis in relative space relations), especially for non-consolidated frontier areas, given that the deforestation process is known to spread from the road network. The results obtained confirm this hypothesis. For the cells connected to the network, the indices were, in average, approximately than 30% higher when using GPMs that take into account the relative space relations (open network criterion). We have selected five representative cells for discussion, shown in Figure 6 below.

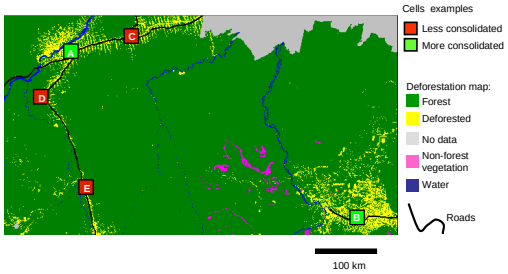

Figure 6. Selected five cells for results presentation.

Figure 7 (see Annex) presents the neighbors of these five selected cells using the two alternative criteria for the GPM construction: local adjacency and open network. We have used indicator functions to build the GPM. The parameters used for the Open Network Criterion were: d_{max} = 20 km ; p_{max} = 50 km. Table 1 presents the Local Moran index values (and corresponding significance) computed for both criteria.

Table 1 – Local Moran index comparison for selected cells.

Case	Classifi	Criteria: Local Adjacency		criteria: Open Network connection	
		L. Moran	Sign (%)	L. Moran	Sign (%)
A	Consol	13,21303	100	11,0170	100
B	Consol	22,26286	100	27,23771	100
C	Inter	4,217319	100	6,002659	100
D	Inter	0,34357	92	0,60697	97
E	New frontier	0,07419	76	0,52020	93

The results confirm, in general, our hypothesis. In consolidated areas, network effects are less important because the local adjacency neighborhood is able to capture the nature of the territorial dynamics. However, going to new frontier areas, the results obtained by the network connection are significantly higher than the ones obtained by the local adjacency neighborhood relations. We intend to continue to study the implications of alternative neighborhood structures in spatial analysis techniques, specially aiming at understanding and modeling the land use and land cover change process in the Amazon.

5 Related Work

Couclelis [1] proposes the notion of *proximal space*, which aims to combine the concepts of absolute space (location) and relative space (situation). To capture relations in proximal space, Couclelis proposes the notion of a *relational map* [1]. Given a set of spatial objects O where each object o_i is associated to a location l_i, a relational map R_i is the set of all locations that influence location l_i. The set of all relational maps for all spatial objects is called the *metarelational map*. The *geo-algebra* proposed by Takeyama and Couclelis [9] uses the metarelational map to extend traditional map algebra operations to operate over the proximal space, and thus captures spatial relations that act at a distance.

The formalism for *geo-algebra* defines a map M as a function $M: L \rightarrow V$ defined as $m=\{(l, m(l)), \forall l \in L\}$, where L represents the set of all locations and V the set of all values associated to these locations. The *geo-algebra* operations can be then defined as follows:

- Let ϕ be an operation over a set of numerical values, such as mean or maximum value;
- Let the *relational map* R_l be defined as the set of all locations that influence a location l;
- Let the set of values V_l by defined by the product of the relational map R_l and the map M ($V_l = M \otimes R$), comprising the values of all locations that influence l;
- Applying the operation ϕ over all the sets V_l generates a new map, as depicted in the equation:

$$M^{new} = \phi\ (M \otimes R_l), \forall\ l \in L. \qquad (10)$$

The *geo-algebra* of Takeyama and Couclelis [9] can be expressed by operations that use the GPM to compute the results for each cell, as follows:

- Let O be a set of spatial objects, each characterized by a location l and a value v associated to each location. Then it follows that S is equivalent to $M: L \rightarrow V$, as defined before;
- Let ϕ be an operation over a set of numerical values, as above;
- Let W be a generalized proximity matrix where each w_{ij} is either one (1) or zero (0), indicating the presence or absence of a relation between the locations l_i and l_j. It follows that each line i of W contains the same information as the relational map R_i;
- For each location l_i, let the set of values V_i be computed as the product of the weights w_{ij} and the values $M(l_j)$, $\forall\ l_j \in L$. In this case, this set of values is the same as the one produced by applying the relational map R_l to the map M;
- The geo-algebra operations can be defined by the application of the operation ϕ to all sets V_i in the map:

$$M(i)^{new} = \phi(\{V_i\}), \text{ where } V_i = \{(w_{ij} * M(l_j))\}, \forall\ l_j \in L. \qquad (11)$$

Therefore, the *geo-algebra* of Takeyama and Couclelis [9] can be obtained by a suitable choice of a GPM and by defining all map algebra functions to be calculated using the GPM. Therefore, it follows that the GPM is a generalization of Couclelis' notion of proximal space. Additionally, Couclelis [1] does not discuss techniques for computing the metarelational map, and does not indicate how absolute and relative space should be combined to compute neighborhood relations.

O'Sullivan [10] proposes a graph-cellular automaton model (or *graph-CA* for short) for the representation of proximal space. A graph-CA extends the basic CA model by using a directed graph G. Each cell c_i of the CA is associated to a vertex v_i of G and each edge of the graph represents a relationship between two cells c_i and c_j. Applications of graph-CAs are presented in O'Sullivan. [11].

The *graph-CA* model can be expressed using the GPM model when a CA uses the GPM to express its neighborhood relations. Recall that a CA can be defined by a tuple (X,S,N,f) in which:

(i) $X \subset Z^2$ is the celular space;
(ii) S is the finite set of possible states;
(iii) $N(x) = \{x_1, \ldots, x_k\}$, is set of cells that are in the neighborhood of a cell $x \in X$.
(iv) $f: S^k \rightarrow S$ is the transition function defined as $S(x,t)=f(S(x_1,t), \ldots, S(x_k,t))$, $\forall\ x_k \in N(x)$, where $S(x_i,t)$ is the state of the CA in position x_i in time t.

A graph-CA is a relaxation of a conventional CA where cell neighborhoods need not be identical, nor local. The relations are defined by a directed graph G, composed of a set of vertices V and edges E, where each cell c_i of the CA is associated to a vertex v_i and each edge of the graph represents a relationship between two cells c_i and c_j. This relationship can be expressed conveniently in a GPM, which contains for each w_{ij} a measure of the relationship between cells i and j. Therefore, any CA whose neighborhood relations are expressed by a GPM will support the graph-CA paradigm.

Therefore, the GPM supports both geo-algebra and graph-CA models of proximal space, and it is more

general than these two definitions. Since it is also a useful tool for computing spatial statistics metrics, the GPM is a convenient and generic way of expressing spatial relations.

6 Conclusions

In today's globalized world, where flows of resources and information are becoming increasingly important, spatial information systems need to incorporate flexible definitions of space. The generalized proximity matrix (GPM), a concept introduced in this paper, is able to combine neighbourhood criteria based both absolute and relative space definitions this allowing to combine local actions with action-at-a-distance. In this paper, we indicate how the matrix can be calculated, considering different types of network configurations. We have also presented a case study where the GPM has been shown to capture spatial relationships that we not detected by considering only local adjacency, by including network connections.

We have implemented the concepts described in this paper using the Terralib environment, an open source GIS library available at http://www.terralib.org[12]. Terralib includes a set of classes to create generalized proximity matrices, allowing the selection and combination of different criteria for construction (e.g., local adjacency and/or network connection), weighing (e.g., inverse of distance and/or minimum path) and slicing (e.g., distance zones and/or adjacency order). Once constructed, tools for spatial analysis or dynamic modeling (e.g., generalized cellular automata) available in TerraLib can be applied using the GPM. Multiple proximity matrices can also be used for different attributes, providing a rich environment for analyzing similarities and dissimilarities, and for exploring complex spatial processes.

References

1. Couclelis, H., *From Cellular Automata to Urban Models: New Principles for Model Development and Implementation*. Environment and Planning B: Planning and Design, 1997. **24**: p. 165-174.
2. Schuurman, N., *Critical GIS: Theorizing an emerging science.* Cartographica, 1999. **36**(4): p. 1-108.
3. Soja, E.W., *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. 1989, London: Verso.
4. Harvey, D., *The Condition of Postmodernity*. 1989, London: Basil Blackwell.
5. Castells, M., *A Sociedade em Rede*. 1999, São Paulo: Paz e Terra.
6. Bailey, T. and A. Gattrel, *Spatial Data Analysis by Example*. 1995, London: Longman.
7. Anselin, L., *Interactive techniques and Exploratory Spatial Data Analysis*, in *Geographical Information Systems: principles, techniques, management and applications*, P. Longley, et al., Editors. 1999, Geoinformation International: Cambridge
8. Anselin, L., *Local indicators of spatial association - LISA.* Geographical Analysis, 1995. **27**: p. 91-115.
9. Takeyama, M. and H. Couclelis, *Map Dynamics: Integrating Cellular Automata and GIS through Geo-Algebra.* International Journal of Geographical Information Systems, 1997. **11**(1): p. 73-91.
10. O'Sullivan, D., *Graph-cellular automata: a generalised discrete urban and regional model.* Environment and Planning B: Planning and Design, 2001. **28**: p. 687-705.
11. O'Sullivan, D., *Exploring spatial process dynamics using irregular graph-based cellular automaton models.* Geographical Analysis, 2001. **33**: p. 1-18.
12. Câmara, G., et al. *TerraLib: Technology in Support of GIS Innovation*. in *II Workshop Brasileiro de Geoinformática, GeoInfo2000*. 2000. São Paulo.

ANNEX

Figure 7. Neighborhood relations identified by alternative criteria for the five selected cells.