

SPATIAL METAPHORS FOR VISUALIZING VERY LARGE DATA ARCHIVES

André Skupin

**National Center for Geographic Information and Analysis
Department of Geography, 105 Wilkeson Quadrangle
State University of New York at Buffalo
Buffalo, NY 14261
voice: (716) 645-2722 ext. 32
email: skupin@geog.buffalo.edu**

and

Barbara P. Battenfield

**Department of Geography, Campus Box 260
University of Colorado, Boulder, CO 80309
voice: (303) 492-3618
email: babs@colorado.edu**

As people try to cope with massive amounts of electronic information, browsing through very large archives can frustrate and eventually impede the retrieval of information. The paper presents an approach to search catalogs of very large electronic archives, that combines computer visualization with the descriptive and analytical power of geography. This method is called “spatialization”, referring to the application of spatial (and visual) metaphors for organizing large volumes of information that are not necessarily spatial in nature.

The power of the metaphor follows from many commonly accepted geographic principles. One of these is Tobler’s Law, that items closer together are more similar than items located farther apart, which is useful for categorizing ‘regions’ in the catalog. Another is the concept of scale-dependence, related to the details which emerge as one observes a geographic landscape more closely. A similar hierarchy can be apparent in archive catalogs, as a person refines a search. Other principles may also apply, for example intervisibility analysis, applicable to cross-referencing various portions of the catalog using a geographic metaphor of “line-of-sight”.

Previous attempts to apply spatial and particularly geographic metaphors to large archives have only partially succeeded. This is due in some part to the technique utilized to ‘locate’ archive items within the confines of the metaphoric ‘space’. We apply multidimensional scaling to descriptive keywords, establishing a numeric coordinate system whose properties can support the principles

identified above. We implement the metaphor of digital terrain representation to a catalog of roughly 100 news stories. This is not a truly large archive, however it demonstrates a proof of concept, and provides a means to identify potential limitations and directions for future research. The paper presents specific details on constructing this spatialization, discusses various problems encountered, and introduces ways in which geospatial information technology may be applied in the process.

INTRODUCTION

Data visualization has emerged as a popular response to the volume of information presented to most people in the course of each day. To assist in comprehending complex data patterns, computer-based visualizations capitalize on the innate human acuity to process visual information efficiently. In the words of McCormick (1987), they allow us to see the unseen. The more successful visualization tools are based on relatively simple metaphors (such as the desktop or layered map overlay) which rely on viewers' familiarity with everyday objects to encourage intuitive manipulations of the information presented.

This paper presents a technical explanation for creating visualizations based on spatial or geographic metaphors, what is referred to as "spatialization". People are accustomed to many metaphors of space, including such aspects as bigger items tend to be more important; higher elevations tend to be more prominent in a landscape. Aspects of things being close together or far apart are readily associated with similarity of characteristics. The argument predicated this paper is that spatial metaphors may be applied to clarify and organize large collections of information. The content of such collections need not be geographic, as demonstrated by earlier applications of spatialization (Erickson, 1993; Chalmers, 1993).

We utilize a case study using news articles, and publish our method here in detail.

NEWSPAPER INFORMATION SPACE

Information spaces come in many forms. Each is characterized by a certain structure and content. For instance, newspapers present information in chunks in the form of articles, which can be more or less related. Although sequentiality plays some role in the structuring of newspaper contents, it is far less important than for instance in a novel. One can quite easily, and without information loss, jump back and forth between newspaper columns. It is possible to read a newspaper from the last page to the first. Much of this is due to the fact that articles are, relative to the whole paper, small and often only marginally related

to each other. Their commonalities are, however, important in determining the positioning of articles within categorical sections, e.g. “national”, “international”, “technology”, “entertainment”.

Given a newspaper article’s unit size and coherence, it is no surprise that queries of computer-based news services are usually based on the article as the most useful meaning-bearing unit. Our research example addresses computer-based information retrieval services. Working with newspaper information spaces, we adopted the article as the basis for the spatialization. The type of information space one is dealing with may dictate other base units in other case studies. For instance, single web pages are a good unit of meaning for spatializing World Wide Web information spaces. Book chapters or journal articles provide reasonable units of meaning for a digital library of textual materials. And so on.

The input data to the spatialization exercise reported here were derived from two editions of the New York Times, Section A (World News), dated November 7 and November 8 1996. These editions went to press a few days after the assassination of the Israeli Premier Yitzhak Rabin, and during the high visibility period following a rape accusation against American military personnel on Okinawa. This two-day period was chosen precisely for the very focused news stories. One quarter of the articles dealt with aspects of these topics. There were articles about the assassin, the funeral, and the mourning among Jews around the world, as well as background reports about the Middle East peace process. These two were primary and secondary dominating themes in the information space. We proceeded from the premise that a robust spatialization method would preserve this dominance in the statistical and cartographic solutions.

About fifty articles were contained in each news edition, with a total of 96 articles in total. The task was to represent all those articles in a spatialized form, based solely on the information content. The following describes the steps leading up to the creation of a three-dimensional model.

CREATION OF A TWO-DIMENSIONAL CONFIGURATION

Our specific approach arose out of three major considerations:

- a) Similarities in the news articles should relate to proximities in the two-dimensional representation. In accordance with Tobler’s premise that in geographical space things that are closer together tend to be more similar, articles of similar content should be positioned closely in the spatialization.
- b) Only the textual content of the articles should be used to construct the spatialization.

- c) The construction procedure should not be arbitrary. It should have a theoretical basis, should be statistically robust, and reproducible by independent efforts.

We utilized accepted methods of text retrieval to process the article content. One such method is based on the vector-space model (Salton, 1989). One computes similarities between textual units based on frequencies of shared keywords. The principles of vector-space modeling are commonly applied in verbal search engines, for instance on the World Wide Web. In these engines a query containing a certain number of terms is compared either to a list of keywords associated with each article or to the whole content of the article.

Keyword assignment

In accordance with the vector-space modeling approach, we assigned a number of keywords to each article. Two methods of keyword extraction may be utilized. They can be (a) extracted explicitly from the text or (b) assigned to articles based on some pre-defined index system. Both methods were applied in this study to determine the impact of extraction method on the spatialized solution. The number of extracted keywords also makes a difference. Preferably, all articles should have the same number of keywords. However, some articles are so short or so general in content that no more than five keywords can be derived. Other articles are quite long or complex so that fifteen or more keywords were extracted. The effects of these differences in methods of extraction and in the number of extracted keywords will be illustrated later in the paper.

Keywords were assigned manually due to a lack of appropriate software and in order to facilitate better understanding of the involved problems. To insure that this process maintained a degree of objectivity, it was repeated independently by several readers, and compared. Results of the keyword assignment proved highly consistent. Table 1 illustrates the selection of keywords for five sample news articles. In the Table, the Article-ID is coded as follows: section a, page number/column number. Multiple keywords were assigned to each article.

Table 1. Assignment of Keywords to Each Article.

Article-#	Article-ID	Term1	Term2	Term3	Term4
...
5	a1/5	jerusalem	assassination	middle east	israel
6	a1/6	king hussein	funeral	rabin	israel
7	a3/1	south america	ecuador	energy	crisis
8	a3/2	south america	colombia	hurtado	assassination
9	a4/1	japan	commercial	advertisement	models
			

Vector-space model

The vector-space model is generated from the set of keywords assigned to the New York Times articles. All the distinct keyword terms are put into the major vector:

$$T = [t_1, t_2, \dots, t_i] = ["AIDS", "advertisement", \dots, "Rabin", \dots, "Zyuganov"]$$

Each article is then matched with this vector to create a Term-Article Matrix (Table 2). In our case a Boolean (1-0) format was chosen, indicating whether a certain keyword is contained in an article. One may also attach weights to each entry to indicate such factors as importance of a keyword, frequency within a document, etc.

Each column in the table represents a distinct vector of keyword matches generated for each article. In the next step, each article vector is related to all other article vectors. The goal is to find a numerical expression describing similarities between articles. There are many possible coefficients to choose from. Some produce similarities, others dissimilarities. We want to express similarity through proximity, that is, high similarity will translate to close (Euclidean) proximity. Using the standard formula for Euclidean distance, articles having many keywords in common will have many terms cancelling (1-1=0). The magnitude of dissimilarity is the square root of the total number of non-similar keywords. Table 3 shows a part of the dissimilarity matrix resulting from the vector comparisons. Along the diagonal, one sees that each article is perfectly similar to itself.

Table 2. Term-Article Matrix.

	Article 1	Article 2	Article 3	Article 4
Keyword 1	0	0	1	0
Keyword 2		0	0	0
Keyword 3	0	0	0	0
Keyword 4	0	0	0	0
Keyword 5	0	0	0	0
Keyword 6	0	0	1	0
Keyword 7	0		1	1

Table 3. Dissimilarity Matrix.

	Article 1	Article 2	Article 3	Article 4
Article 1	0	4.123	4.690	4.242
Article 2	4.123	0	4.582	3.872
Article 3	4.690	4.582	0	4.242
Article 4	4.242	3.872	4.242	0

Multidimensional scaling

Multidimensional scaling (MDS) has been a much discussed procedure throughout the social sciences (Golledge and Rushton, 1972). In the past, MDS has been applied to map behavioral space (Buttenfield, 1986), travel space (Wolfe, 1978), and disciplinary space (Goodchild and Janelle 1988). The majority of MDS applications in geography have focused on analytical tasks or to comprehend complex relationships of the subject of study. In this context, the focus of the MDS research is on finding geometric dimensions that best represent the underlying structure of a data set. Shepard (1966) devised procedures to determine whether the optimal representation should be visualized in two, three, four, or more dimensions.

Our focus is on the support of fast retrieval and interactivity for large data bases, rather than on the underlying structure of the information space as an entity. This research is intended to serve as a proof-of-concept for applying spatial metaphors to non-geographic information spaces. In the example presented here, we restrict the solution to two spatial dimensions.

It lies beyond the scope of this paper to explain the mathematical and statistical background of MDS. Even Kruskal's classical overview of MDS does not cover its calculations, since "even the simplest versions are virtually never performed without a computer" (Kruskal 1978, p.15). For more in-depth discussion refer to Young et al. (1987), for some theoretical aspects of dimensionality to Jacoby (1991) and for a bibliographical overview to Coxon & Jones (1983). An excellent presentation of the geometry is given by Tobler (1976).

Input to MDS is the dissimilarity matrix (Table 3). The respective module of SPSS was used to run the MDS procedure. Its main result is a configuration with two-dimensional coordinate pairs for each article (Table 4).

Table 4. Two-dimensional Coordinate Configuration.

Article-#	X	Y
1	1.313	-0.661
2	-1.790	-0.129
3	0.927	-0.573
4	0.774	-1.092
5	-1.821	-0.415

VISUAL DISPLAYS OF THE SPATIALIZATION

Earlier it was mentioned that the visualization of spatialized data is one goal of this work. The use of GIS software becomes relevant as soon as the solutions can be configured in two-dimensional coordinate systems. Two types of visual displays are discussed below. The first is based on point scatter graphs, and the second simulates a terrain representation.

Point Visualization

The two-dimensional coordinates (Table 4) were input to ArcView in the form of event files (ESRI terminology) which made them readily available for visualization. The coordinate files were linked to the original keyword file (refer back to Table 1) through common identifiers, in straightforward GIS manner. This made it possible to control the quality and validity of the spatialized solution by comparing highlighted items in the keyword table with the respective points in the scatter plot. Examples for point-based visualization can be seen in Figures 1, 2, and 3.

Figure 1 compares solutions for the two methods of keyword extraction. As mentioned above, those methods have an effect on the geometry of the spatialized solution. Keywords may be extracted in different ways. They can be (a) extracted directly and explicitly from the text or (b) assigned to the respective text based on some index system. In our case that distinction proves to be most important. Figure 1 is an ArcView-based illustration of how these differences influence the spatialization results. In both windows the same articles are highlighted, all of which cover a variety of foreign issues from a U.S. perspective. In the "Euclidean Concordant" window, keywords were extracted directly. Here, the solution displays highlighted articles as two separate clusters. In the "Euclidean Indexed" window, keywords were built up from a pre-determined index. Here, the highlighted articles form one distinct cluster.

This analysis of articles is taken from too small a sample to determine which method of keyword extraction is "best", as that will depend on the data and on the goal of the spatialization. For the proof-of-concept, it is important to note that such differences exist, and to be alert when extracting keywords.

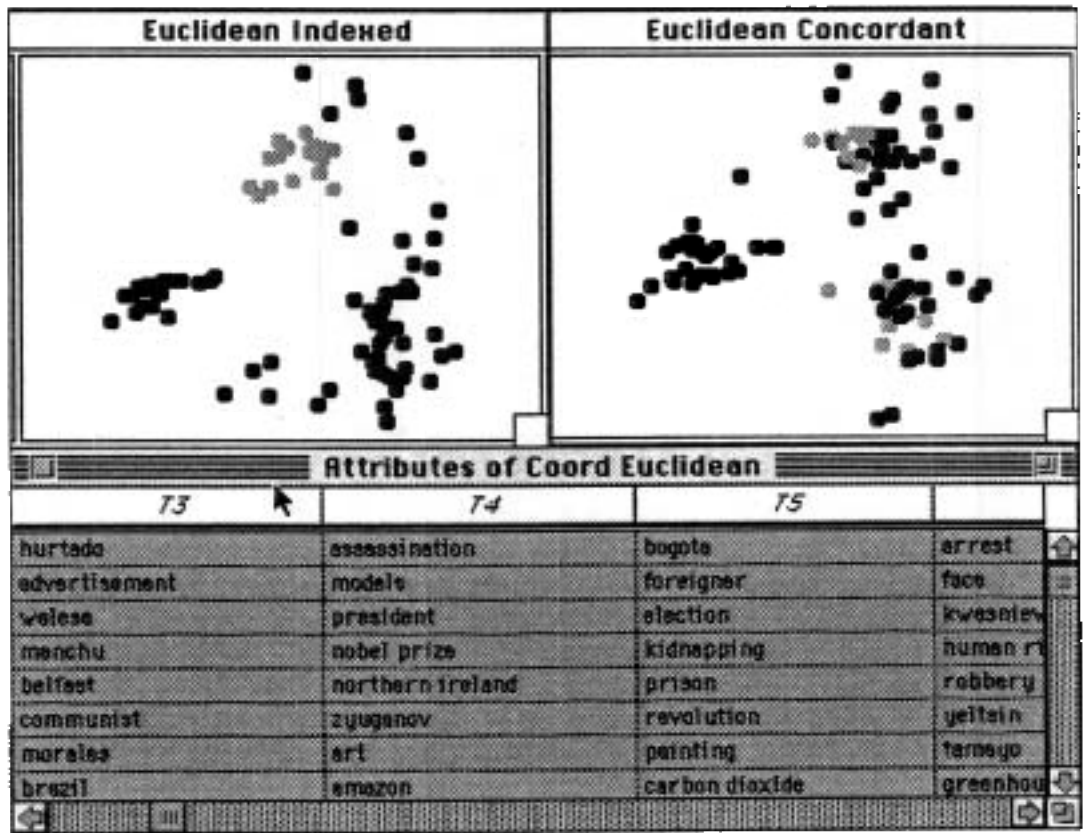


Figure 1. Spatialized Solutions Drawn from Indexed and Concordant Keywords.

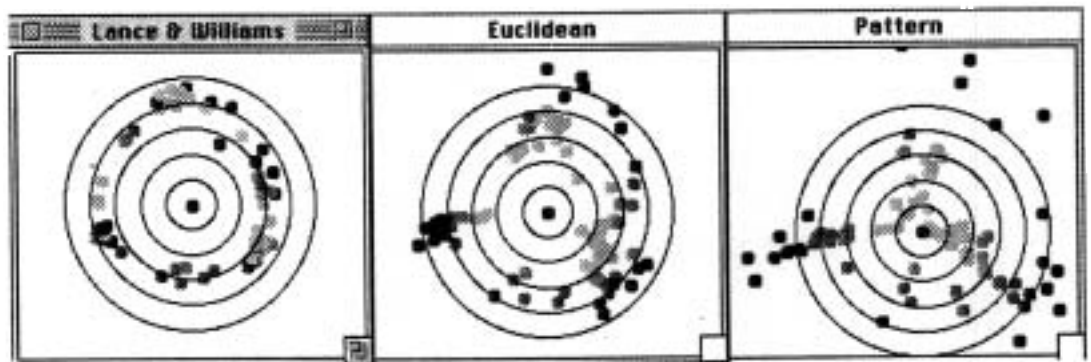


Figure 2. Point Visualization of the Number of Keywords Per Article.

A related problem is posed by the number of keywords assigned to articles. That number affects the vector-space model and thus the spatialized solution. The effects of the number of keywords are illustrated in Figure 2. The total number of keywords for each article is visualized through a gray scale, with light tones for high numbers and dark tones for low numbers. The window titles

“Lance & Williams”, “Euclidean”, and “Pattern” refer to the respective options of the SPSS software. A number of concentric rings is added, with the origin of the MDS configuration [0,0] as the center. This serves as a reminder that all three configurations are shown at the same “scale”, since the rings are plotted at equal distances from the coordinate origin in the coordinate space of the configurations.

There is a progression from a configuration in which articles are placed in a wide stretch from near the coordinate origin into the two-dimensional space (“Pattern”) to a configuration with little variation in the distance of articles from the origin (“Lance & Williams”). For the “Pattern” visualization there is a clear relationship between the number of keywords and the distance of an article from the center of the configuration. For the “Euclidean” measure this is, to a lesser degree, also true. In the case of the “Lance & Williams” measure we see instead a ring structure that has been acknowledged in many MDS solutions.

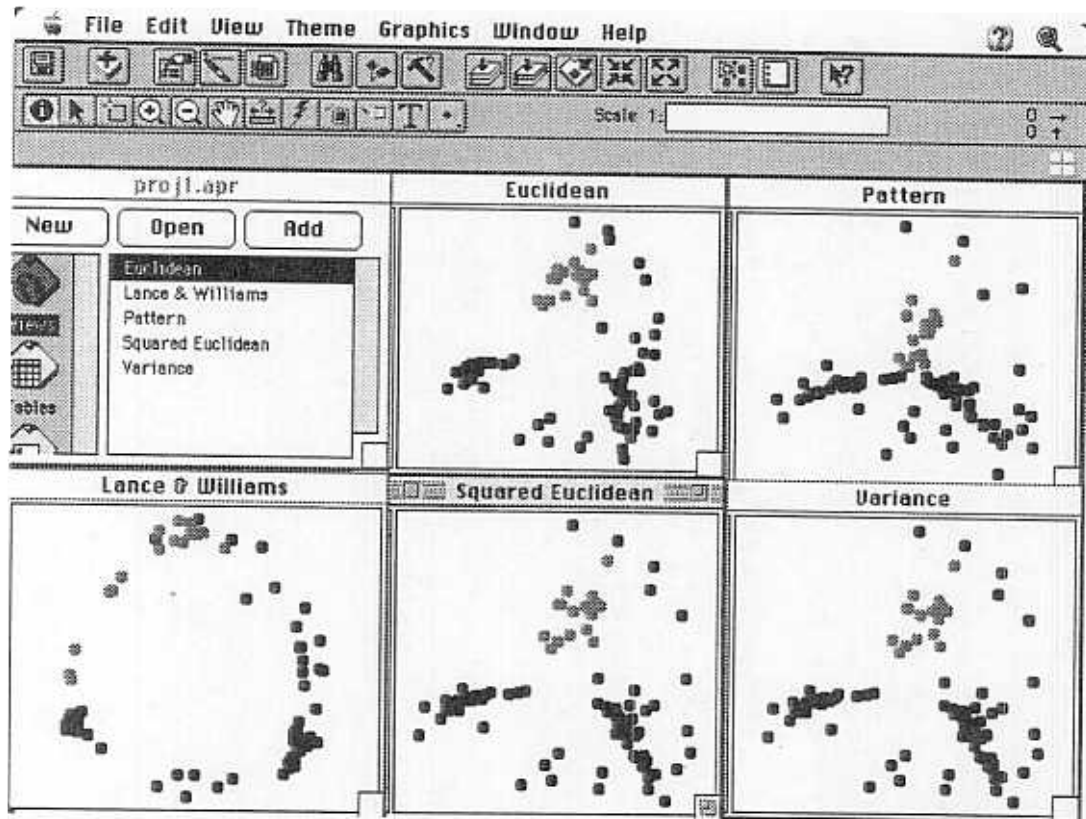


Figure 3. Application of Five Different Proximity Measures to the Same Data Set.

The principles and formulas of similarity/dissimilarity measures can differ quite dramatically and this has a major impact on the spatialized solution. Unfortunately, few objective criteria are established for choosing among them.

This forms a common criticism of the vector-space model. Figure 3 shows how differing similarity measures applied to the same set of articles and keywords can produce very different configurations.

Landscape Visualization

To generate a terrain or landscape representation of the data, some numeric attribute value must be attached to each article. One option is to compute an attribute value based on a list of the absolute frequencies with which each keyword appears in all the articles. For each article the sum of absolute frequencies of all those keywords is computed, which are actually associated with that article. Imagine that an article contains three keywords, and these keywords appear in six, four, and seven articles, overall. Then the z value would be computed as seventeen.

If the creation of a landscape depiction of the information space is one goal of the spatialization, then appropriate interpolation procedures must be chosen. There is a wealth of experience about geographic surface interpolation within the GIS community. It remains to be seen whether analogies to surfaces representing large non-geographic databases can be made and translated into reasonable interpolation rules. Based on the x-y-z point data a landscape surface was interpolated via universal kriging using ARC/INFO. The surface is displayed as a color composite that includes hillshading. The landscape illustration is not very legible in black and white, and therefore is not included in the proceedings volume. However it will be shown at the conference presentation.

NEW PROBLEMS IN A NEW REALM

We have presented a technical explanation of a reproducible method for constructing coordinate-based representations of information archives. We have shown two visualization methods by which to display results. We have demonstrated several parameters that may be varied in generating spatialized solutions. The method of keyword extraction has an effect on the geometry of the spatialized solution, as does the number of keywords extracted per article. The method of computing similarities between vectors can also modify the solution. We do not view these as limitations of the method, but rather as one might view parameters in any modeling situation, as characteristics that can be adjusted to fit particular applications.

The employment of geographic/cartographic notions, expertise and technology for the visualization of non-geographic information spaces is a relatively new concept. To our knowledge there has been as yet no subject testing to determine the appropriateness of such methods for visualization. However there is enormous potential in database and visualization functionality of GIS technology to organize and navigate through information spaces containing non-geographic information.

REFERENCES

- Buttenfield, B.P., 1986. Comparing Distortion on Sketch Maps and MDS Configurations. *Professional Geographer*, 38(3):238-246.
- Chalmers, M., 1993. Using a Landscape Metaphor to Represent a Corpus of Documents. In: Frank, A.U. and Campari, I. (Eds.) *Spatial Information Theory: A Theoretical Basis for GIS*. Springer-Verlag, Lecture Notes in Computer Science No. 716:377-390.
- Coxon, A.P.M, Jones, C.L., 1983. Multidimensional Scaling. In: McKay, D. et al. (Eds.) *Data Analysis and the Social Sciences*. London: Frances Pinter.
- Erickson, T., 1993. From Interface to Interplace: The Spatial Environment as a Medium for Interaction. In: Frank, A.U. and Campari, I. (Eds.) *Spatial Information Theory: A Theoretical Basis for GIS*. Springer-Verlag, Lecture Notes in Computer Science, No. 716:391-404.
- Goodchild, M.F., Janelle, D.G., 1988. Specialization in the structure and organization of geography. *Annals of the Association of American Geographers*, 78:1-28.
- Golledge, R. G. and Rushton, G., 1972. *Multidimensional Scaling: Review and Geographic Applications*. Association of American Geographers Commission on College Geography, Technical Paper No.10. Washington, D.C.
- Jacoby, W.G., 1991. *Data Theory and Dimensional Analysis*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-078. Newbury Park, CA: Sage Publications.
- Kruskal, J.B., Wish, M., 1978. *Multidimensional Scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011. Beverly Hills and London: Sage Publications.
- McCormick, B.H., DeFanti, T.A., and Brown, M.D. (Eds.), 1987. *Visualization in Scientific Computing*. SIGGRAPH Computer Graphics Newsletter, 21(6).
- Salton, G., 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company.
- Shepard, R. N., 1966. Metric Structures in Ordinal Data. *Journal of Mathematical Psychology*, 3:287-315.
- Tobler, W.R., 1976. The Geometry of Mental Maps. In: Golledge, R. G. and Rushton, G. (Eds.) *Spatial Choice and Spatial Behavior*. Columbus, Ohio: Ohio State University Press: 69-82.
- Wolfe, R.A., 1978. *Refinement and Applications of Time-Space Mapping Techniques*. University of Toronto/ York University Joint Program in Transportation, Research Report No. 46.
- Young, F.W., Hamer R.M., 1987. *Multidimensional Scaling: History, Theory, and Applications*. Hillsdale, N.J.: Lawrence Erlbaum Associates.