Distance imprecision and error in spatial decision support systems

Marc P Armstrong
University of Iowa

Posted with permission of the author


Hosted by Iowa Research Online. For more information please contact: lib-ir@uiowa.edu.
INTERNATIONAL GEOGRAPHIC INFORMATION SYSTEMS (IGIS) SYMPOSIUM

The Research Agenda

November 15-18, 1987
Arlington, Virginia

Proceedings

Volume II

Technical Issues and the Research Agenda

Dr. Robert T. Aangeenbrug, Co-Editor
Yale M. Schiffman, Co-Editor

Convened by the
Association of American Geographers

In Cooperation with
National Aeronautics and Space Administration
Bureau of Land Management
U.S. Army Engineer Topographic Laboratories
National Oceanic and Atmospheric Administration
U.S. Geological Survey
Defense Mapping Agency
Bureau of the Census
Urban and Regional Information Systems Association
American Congress on Surveying and Mapping
World Computer Graphics Association
American Society for Photogrammetry and Remote Sensing
Marine Technology Society
Centre for Earth Resource Management
National Governor's Association
Society of American Foresters
National Agricultural Chemicals Association
Spot Image Corporation
Earth Observation Satellite Company
Distance Imprecision and Error in Spatial Decision Support Systems

Marc P. Armstrong
Departments of Geography and Computer Science
316 Jessup Hall
University of Iowa
Iowa City, IA 52242
BLAMMGPD@UIAMVS.BITNET

ABSTRACT

Locational models are being used with increasing frequency in geographic information systems. Many such models use distance measurements in calculating solutions. A number of commonly used distance measurement procedures are used to examine whether results are influenced by the various estimates. In one set of analyses, the objective function increases with the scale factor of the source map. Additionally, the geometrical configurations yielded by the analyses vary among the distance estimation procedures.

1.0 INTRODUCTION

Geographic information systems (GIS) use sophisticated data capture and display functions to provide assistance in solving a range of social and environmental problems. At the same time, locational analysis methods continually improve, and system designers and academic researchers are now integrating these analysis procedures into GIS toolboxes (Densham and Armstrong, 1987; Elmes and Harris, 1986; Lupien, Moreland, and Dangermond, 1987). This melding of analytical GIS and locational modeling capabilities has led to the formulation of spatial decision support systems (DSS), in which alternative locational decisions can be explored and evaluated (Armstrong, Densham, and Rushton, 1986; Davis and Grant, 1987).

As GIS and DSS technology is brought to bear on important, and potentially threatening, environmental and social issues, the results and recommendations stemming from the application of the technology will be carefully examined. Obtaining insight into the nature of imprecision and error is important in this context, because if a DSS is to be used effectively, the effects of error and imprecision on results must be recognized.
This paper examines one facet of error. The specific goal is to evaluate the role of commonly used distance estimates in calculating the objective functions and resulting geometrical configurations of p-median analysis procedures used to support spatial decision-making.

2.0 BACKGROUND AND PROBLEM ELABORATION

2.1 Concepts of Distance

Distance has long been cited as a fundamental concern of quantitative geography (Nystuen, 1963). Many researchers employ distances (e.g. miles) based on the general Minkowski metric (Kuiper, 1986). Others have used alternative distance measures such as time-distance, and cost-distance, both of which are functional measures of effort (Falk and Abler, 1980). While these functional distances are important, metric distances continue to be used because they are convenient to use and because other measures of distance often can be derived from them.

2.2 Spatial Data Organization and Distance Computation

Although geographic information systems store data in various ways, many define objects using a vector approach (Peuquet, 1984; White, 1984). When data are vector coded, objects such as road networks are often defined by chains of points with associated topological linkages. Network distances are computed in vector databases by summing scaled interpoint distances within each chain. Given measurement costs, analysts may wish to rely on existing data to compute distances. For example, the U.S. Census Bureau has developed digital data products to support its operations (e.g. DIME), and the USGS is actively involved in encoding and disseminating digital data. These and other digital sources will play increasingly prominent roles in spatial modeling during the coming decade.

2.3 Error in Digital Representations of Lines

Peucker (1975) has outlined a general theory of cartographic lines. The theory is especially pertinent to the treatment of line complexity. Chrisman (1982) has developed a preliminary theory for assessing
cartographic error and its measurement. Other researchers (e.g. Muller, 1987) have addressed specific dimensions of error and have identified a series of error sources. These sources can be placed into the following categories:

**Sampling density.** The well-known (e.g. Perkal, 1966; Buttenfield, 1985) Steinhaus paradox shows that as smaller sampling intervals are used to measure a sinuous feature, it apparently increases in length. Sinuosity which was missed in previous intervals is now detectable, and is added to the previous length estimate. During digitizing, decisions are made about data encoding precision. These decisions affect distance calculations.

**Source material scale.** The scale at which a map is compiled can influence length measurement because generalization, in the form of decreased line complexity, increases as map scales become smaller (Gardiner, 1982; Maling, 1968).

**Source material error.** Dueker (1975) has described error types that occur on source maps. If the source contains error, its digital product will, at best, replicate those errors with a high degree of fidelity. It is likely, however, that additional error will be introduced during data capture.

**Error in data capture.** When documents are captured in digital vector form, encoding errors occur (Dueker, 1975). Jenks (1981) has identified errors that occur during manual encoding of irregular linear features.

**Error in data storage and processing.** Chrisman (1984) argues for using good judgment in selecting the resolution for coordinate storage, and for the use of absolute, rather than incremental coordinates. Chrisman also describes other issues that must be weighed (e.g. real vs. integer, and decimal precision).

**Topological error.** Ginsburgh and Hansen (1974) describe three common error types in studies using road networks. Their third type deals with topological problems arising from incomplete specification of internodal linkages. In their framework the topology of a network is verified by reciprocal enumeration.

**Aggregation error.** Hillsman and Rhoda (1978), Goodchild (1979), and Casillas (1983) have shown that aggregating
demand (e.g. households) to a single proximal point misrepresents the actual underlying distribution of demand and introduces error into analyses.

Also note that in some instances, other factors contribute to measurement error (e.g. document instability, projection). Each of the above factors can act alone, but error is often attributable to several factors acting in concert.

2.4 Distance Estimation

Many procedures used by decision support systems optimize distance-based objective functions. Researchers have suspected that simple interpoint distances are not always appropriate. As a result, locational analysis software (e.g. Goodchild and Noronha, 1983) often allows alternatives to symmetrical interpoint distances (e.g. time, or asymmetrical distance matrices to model one way streets).

Varying interpoint distances by a constant also has been the subject of empirical research because of differences between straight-line and network distances. Nordbeck (1963), for example, found that a constant (1.23) could be applied to straight-line distance to compensate for measurement problems. Love and Morris (1972; 1979) also examined several methods (e.g. Euclidean and spherical) for producing estimates, and compared each to 'actual' distances obtained from road map distance matrices and government agencies. The estimates were also parametrically altered. For example, simple Euclidean distances performed poorly when compared to estimates, such as:

\[ D_{ij} = 1.16 \times (\text{abs}(x_i-x_j) \times 1.60)^{1/1.62}, \]  

where \( i \) and \( j \) represent two locations for coordinate pair \( x \), and 1.16, 1.60, and 1.62 are estimated parameters.

Kolesar (1979) provides a distance approximation for deriving estimated travel times based on distances within U.S. urban areas:

\[ D_{ij} = 1.15 \times \sqrt{(X_i-X_j)^2+(Y_i-Y_j)^2)}, \]

where 1.15 is an empirical constant. The size and form of parameters change among these studies (e.g. 1.23; 1.16; 1.15), and Ginsburgh and Hansen (1974) assert that
values ranging between Euclidean distance, and Euclidean distance \* (1+SQRT(2)) are acceptable. This variability accentuates uncertainty in producing distance estimates, leads to problems with interpretation of results, and can lead to an inability to replicate spatial modeling experiments.

The entire estimation process may be obviated if 'actual' distances in the form of digital networks are available for analyses. Kolesar (1979), for example, suggests that using network models of actual street data to compute distances can yield accurate results. He does, however, note that 'extensive data base and time consuming calculations' are required (Kolesar, 1979). Information of this kind is available in the form of USGS Digital Line Graphs (DLG). These DLG data are an ideal source of reference data, because they are encoded at a high level of precision and are topologically validated (U.S.G.S., 1985).

3.0 METHODS

3.1 Study Area

A series of commonly used distance estimates are used to assess the magnitude of changes which occur in the locational analysis procedures. The site for the research is the Chickamauga 1:100,000 quadrangle in GA, AL, and TN. The site contains parallel ridges, and the road network, in part, reflects the topography; many roads running orthogonal to the ridges tend to be more sinuous than those running parallel to the ridges.

The quadrangle, which is a primary data source in the research, conforms to National Map Accuracy Standards: not more than 10% of points on the map are in error of more than 1/30 inch, measured at the publication scale (Thompson, 1979). This standard, however, applies to 'well-defined' points such as road intersections and benchmarks. Error in the trace of linear features, therefore, may be uncontrolled by the accuracy standard.

3.2 Demand Weights

A sample of 26 demand nodes was drawn from named places on the Chickamauga quadrangle, and if data were available, each node was assigned a population weight from the 1980 Census; for unincorporated places without
available population figures, a weight derived from the mean of small places with known population, was assigned. An additional set of analyses was performed on the nodes with each weight set equal to one. The intent of this second set is to examine purely spatial effects on the outcome of the analyses.

3.3 Distance Calculations

The data sets, or estimation procedures used are:

1) Manually encoded network at 1:100,000;
2) Manually encoded network at 1:250,000;
3) Manually encoded network at 1:500,000;
4) Euclidean internode distance;
5) Manhattan internode distance;
6) Parameterized Euclidean distance (Equation 1).

Manual digitizing. Distances were measured using point sample digitizing on a series of USGS maps: Chickamauga, 1:100,000; Rome, 1:250,000; and the USGS state maps for Georgia and Alabama, 1:500,000.

Euclidean, Manhattan, and Parameterized. The location of each demand node was encoded, and a program calculated distances for each metric.

3.4 Data Analysis

The distributions were processed using a shortest-path algorithm and a vertex substitution method for solving the p-median problem (Tietz and Bart, 1968). For each analysis performed, the following questions were addressed: How does the value of the objective function change? Are different geometrical configurations produced when the distance measures are changed?

4.0 RESULTS

The analyses for the 26 nodes were performed with 10 sites to be assigned facilities. Each node was a site candidate.

4.1 Weighted Analyses

Objective function. The analyses for population-weighted data exhibited consistency across some of the
distance measures (Table 1). For example, the results from the parametric, Manhattan, and the manually digitized networks at 1:100,000 and 1:250,000 have similar objective functions. The results, however, from the Euclidean and 1:500,000 digitized database, reflect large differences in the objective function.

**Table 1. Results of weighted analyses.**

<table>
<thead>
<tr>
<th>DISTANCE METRIC</th>
<th>OBJECTIVE FUNCTION</th>
<th>% OF EUCLIDEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>14,319.87</td>
<td>100</td>
</tr>
<tr>
<td>Manhattan</td>
<td>18,378.99</td>
<td>128</td>
</tr>
<tr>
<td>Parametric</td>
<td>17,134.81</td>
<td>120</td>
</tr>
<tr>
<td>1:100,000</td>
<td>17,678.11</td>
<td>124</td>
</tr>
<tr>
<td>1:250,000</td>
<td>17,629.29</td>
<td>123</td>
</tr>
<tr>
<td>1:500,000</td>
<td>20,463.39</td>
<td>143</td>
</tr>
</tbody>
</table>

Geometrical configuration. The 1:100,000 and 1:250,000 maps have identical configurations. The manually digitized map at 1:500,000, however, yielded a different configuration (Figure 1). This is clearly an inappropriate scale of analysis for the example problem, because of the abstraction of the road network. Additionally, the Euclidean and parametric analyses yielded identical configurations, that were different from the Manhattan or digitized results.

**FIGURE 1. Geometrical configurations from weighted analysis.**
4.2 Unweighted Analyses

Objective function. The unweighted analyses reflect values attributable to distance variation. Although the objective function fluctuates slightly across analyses, it is notable that for the digitized series, it monotonically increases with the scale factor (Table 2). A similar pattern appeared in the weighted analyses (Table 1). Although the 1:100,000 and 1:250,000 analyses are reversed, their values are quite similar, and are much lower than the 1:500,000 results.

<table>
<thead>
<tr>
<th>DISTANCE METRIC</th>
<th>OBJECTIVE FUNCTION</th>
<th>% OF EUCLIDEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean</td>
<td>92.38</td>
<td>100</td>
</tr>
<tr>
<td>Manhattan</td>
<td>120.74</td>
<td>131</td>
</tr>
<tr>
<td>Parametric</td>
<td>111.16</td>
<td>120</td>
</tr>
<tr>
<td>1:100,000</td>
<td>117.46</td>
<td>127</td>
</tr>
<tr>
<td>1:250,000</td>
<td>118.46</td>
<td>128</td>
</tr>
<tr>
<td>1:500,000</td>
<td>129.61</td>
<td>140</td>
</tr>
</tbody>
</table>

Table 2. Results of unweighted analyses.

FIGURE 2. Geometrical configurations from unweighted analysis.
**Geometrical configuration.** The results exhibit a higher amount of instability than the weighted analyses. The Euclidean and parametric results are identical. The configuration, however, is different than that produced by the weighted analyses. The Manhattan metric produces a similar solution; it differs only because one node is assigned to a different center. Each of the digitized maps produced a unique configuration. Figure 2 shows a comparison of the 1:100,000 and 1:500,000 results.

### 5.0 CONCLUSIONS

This research contributes to knowledge about the workings of analysis procedures that will be used to assist decision makers in making judgments about important future issues. The main goal was to examine variations that result when different distance representations are used in locational models. The results indicate that outcomes are influenced by the scale of the source material or distance estimation procedure. Both the value of the objective function and the geometrical configuration of the allocation are altered when different, but commonly employed, distance estimation procedures are used in the analysis. An important effect centers on the extreme generalization of some data sources. Although generalization, and decreased sinuosity, of small scale maps are important in calculating distances, the decrease in distance is offset by another facet of generalization—exclusion of low-order roads.

Future work will focus on use of existing digital data products. Most important is an examination of differences that may arise from the use of simple interpoint representations, digitized networks, and distance calculated from existing spatial databases (e.g. DLG).

### ACKNOWLEDGEMENTS

The assistance of Panos Lolonis in data collection and processing is gratefully acknowledged. Paul J. Densham coded the location-allocation software and provided insight during the course of performing analyses. Jeff Hanson and David Bennet also provided assistance.
REFERENCES


II-32


Goodchild, M.F. and Noronha, V. 1983. Location-Allocation for Small Computers, Monograph 8, Dept. of Geography, University of Iowa, Iowa City, IA.


