1-1-1989

Digitizing and distance estimation errors in spatial analysis: sinuosity and network configuration effects

Marc P Armstrong
University of Iowa

Brian T Dalziel
University of Iowa

PROCEEDINGS OF

INTERNATIONAL GEOGRAPHIC INFORMATION SYSTEMS (IGIS) SYMPOSIUM '89:

Global Systems Science - An Effective Response to Human Needs

March 18 & 19, 1989
Baltimore, Maryland

Sponsored by

ASSOCIATION OF AMERICAN GEOGRAPHERS
AND
THE CENTER FOR EARTH RESOURCE MANAGEMENT
APPLICATIONS, INC.

In cooperation with

U.S. Geological Survey
U.S. Bureau of the Census
Earth Observation Satellite Company

Prepared by

E.H. Pechan and Associates, Inc.

Digitizing and Distance Estimation Errors in Spatial Analysis: Sinuosity and Network Configuration Effects

Marc P. Armstrong
and
Brian T. Dalziel

Department of Geography
316 Jessup Hall
The University of Iowa
Iowa City, IA 52242

ABSTRACT

In this paper, we examine facets of error that occur when parametric and digitizing approaches are used in distance estimation. First, Euclidean, Manhattan and parametric distance metrics are briefly described. Next, the distance estimation procedures are examined and compared to a normative measure in study areas selected on the basis of road network configuration. The relative efficacy of the different estimation procedures in each area is described. We also examine a series of lines with varying amounts of sinuosity to determine variation in distance estimates derived from manual digitizing.

1.0 INTRODUCTION

Geographic information systems employ data management, display and analysis functions to provide assistance in solving a range of social and environmental problems. Mathematical modeling procedures are also being incorporated into the GIS framework to provide additional assistance in spatial problem solving (Armstrong, Densham and Rushton, 1986; Davis and Grant, 1987; Lupien, Moreland and Dangermond, 1987). Many mathematical spatial models depend upon distance measurements in their computation, and several means for providing distance estimates have been developed. Some methods attempt to replicate the form of lines (roads) in digital or analog form, while others either assume away sinuosity in road networks, or attempt to compensate for differences between straight-line and network distance by estimating parameters which are then used to inflate straight-line distances. Variation in distance estimation, however, may affect the results, and hence, the reliability of spatial models used in conjunction with GIS technology. Armstrong (1988) has shown how the use of different distance estimating techniques effects the objective function and optimal configurations generated by the p-median model. Krarup and Pruzan (1980) also evaluate the impacts of distance on locational problems. Given the important uses to which these systems are, and will be, applied, we must acquire knowledge about how these models respond to imprecision and error of the magnitude likely to occur in spatial databases. As a preliminary step, the goal of this paper is to evaluate relationships among distance estimation procedures that are commonly used in spatial models.

2.0 ALTERNATIVE DISTANCE ESTIMATION TECHNIQUES

In this section we review several methods for making distance estimates: the Euclidean metric, the Manhattan metric, the l_k, l_p, $l_{k,p}$, and $l_{k,p,s}$ distance metrics, and the road network model. Most applications identified in the distance estimation literature use mathematical distance estimating functions which are a variant of the Minkowski metric (Kuiper, 1986) :

$$d(a,b) = (| a_1 - b_1 |^p + | a_2 - b_2 |^p)^{1/p}, p \geq 1, \qquad \text{Eq. 1}$$

where d(a,b) is the distance from a to b, a = (a_1,a_2) and b = (b_1,b_2) are points. Special cases of this equation are when p=1, the Manhattan distance obtains, and when p=2, the Euclidean distance obtains.

The Euclidean metric, which yields the shortest distance between two points on a plane, has been widely applied because it is easy to calculate and may be sufficiently accurate for some applications (Laporte, et al., 1985). In most networks, however, distances are not Euclidean -- they are longer, and therefore, the Euclidean metric underestimates distances (Love and Morris, 1972, 1979; Love et al., 1988). This distinction is especially critical in areas where travel is constrained to a rectangular grid. In those cases, for places parallel to one axis, the Manhattan metric equals Euclidean metric distances. But for any other combination of places, the Manhattan metric calculates distances along the grid, rather than along the diagonal between points. As a result, the Manhattan metric overestimates actual distances in many network configurations. In most places road networks are not constrained to a perfect rectangular grid, but fill out an intermediate position on a continuum between rectangular and Euclidean while others still are greater than rectangular (Love and Morris, 1972). As a result, many different functions, known collectively as l_p functions, have been designed to estimate distances across networks. These parametric functions often, but not always, produce estimates which are greater than Euclidean, and less than Manhattan:

$$l_k = d_k(a,b) = k (| a_1 - b_1 |^2 + | a_2 - b_2 |^2)^{1/2} \quad ,k \geq 1 \quad \text{Eq. 2}$$

$$l_p = d_p(a,b) = (| a_1 - b_1 |^p + | a_2 - b_2 |^p)^{1/p} \quad ,p \geq 1 \quad \text{Eq. 3}$$

$$l_{k,p} = d_{k,p}(a,b) = k (| a_1 - b_1 |^p + | a_2 - b_2 |^p)^{1/p} \qquad \text{Eq. 4}$$

$$l_{k,p,s} = d_{k,p,s}(a,b) = k (| a_1 - b_1 |^p + | a_2 - b_2 |^p)^{1/s} \qquad \text{Eq. 5}$$

where k, p and s are parameters.

Nordbeck (1963) and Fildes and Westwood (1978) argue that, for homogeneous road networks, inflating the Euclidean distance with a constant will yield accurate estimates of road network distances (e.g. equation 2). To solve for a lack of road network homogeneity, Fildes and Westwood suggest that small homogeneous regions be identified and that the constant k be optimized for each. They term this the multi-regional model of distance calculation in which the individual k's in each region contribute to the total calculated distance value.

Love and Morris (1972, 1979) measured distances among twelve cities in rural Wisconsin and twelve larger cities from across the nation, and applied Manhattan, Euclidean, and distances derived from Eqs. 2, 3, 4, and 5 to the corresponding points. They argue that Eqs. 4 and 5 estimate distances more accurately than others in their study because they use parameters which "fit" the functions to the study area by minimizing differences between estimated and actual road network distances. Because of the earth's curvature Love and Morris (1972) also compared three spherical distance functions to Eqs. 4 and 5 and conclude that the spherical functions, even for long distances, are inferior. Love and Morris assert that since the differences in values typically calculated for "p" and "s" are small, Eq. 4 can be used in place of Eq. 5. In a similar vein, Berens and Korling (1985) examined West Germany's road network. Although the authors agree that Eqs. 4 and 5 produce estimates with less error than those produced by Eq. 2, they argue that the difference in accuracy between Eqs. 4 and 5 and Eq. 2 is statistically insignificant.

The final distance estimation technique uses a metric in a different way, and represents an alternative to the mathematical estimation approaches described above. Nordbeck (1963) and Francis, McGinnis and White (1983) describe the basic properties of a computerized road network composed of nodes (intersections) and chains (road links) digitized from a source map. Within each chain, a Euclidean estimate is usually used to measure interpoint distances, which are then summed and scaled to produce length estimates. The length of a path on a network, therefore, is calculated by summing chain lengths. For large networks it is helpful, and often necessary, to use shortest path algorithms to find distances between specified nodes. For large areas, data acquisition costs may be high, and thus, the creation of a network may be cost-prohibitive. For small areas, however, networks may yield high accuracy with manageable costs. One source of existing network data is Digital Line Graph (DLG) files produced by the USGS.

3.0 METHODS

In an effort to gauge the stability of different distance estimation procedures, we examined the performance of five functions (Manhattan, Euclidean, and eq. 2, 3, and 4) in five study areas (Table 1) within different USGS 1:100,000 scale maps. In each area, distances between selected places on the road network are estimated using each of the five functions. Each study area has been chosen due to a distinguishing road network characteristic in order to determine, for example, if one function performs better on dense urban networks, and another on rural networks.

Ten nodes were selected within each study area, and each node was digitized to estimate its UTM coordinate. Because digitizing error can seriously affect our results, each node was digitized five times, extreme values were dropped, and the mean was taken from the remaining values to determine each coordinate. Each of the 45 combinations of two nodes is referred to as a "node pair."

Table 1 : Study Areas

Location	Type of Map	Date	Key Characteristic
Estherville, IA	30 x 60 min. quad.	1985	rural grid
Vail, CO	30 x 60 min. quad.	1980	mountainous
Newark, NJ	30 x 60 min. quad.	1984	dense urban grid
Gulfport, MS, LA	30 x 60 min. quad.	1982	coastal
Salton Sea, CA	30 x 60 min. quad.	1985	rural grid

Using the UTM coordinates, each metric was used to estimate the distance between all node pairs. The parameters in Eqs.2,3 and 4 (k and p), were optimized using Root Mean Squared Error (RMS) as the goodness of fit criterion:

$$\text{RMS} = \text{Sqrt} \left(\sum_{i=1}^{n} [\, DE(i) - DA(i) \,]^2 / n \right)$$

where: DE = Estimated Distance
DA = Actual Distance

Actual distances were measured from the five USGS maps. For each node pair, an analog map measurer was used to measure shortest path lengths on the road network between node pairs. The shortest path was determined visually, with "close calls" being solved by measuring likely paths and selecting the shortest.

4.0 RESULTS

We use RMS estimates (Table 2) and regression analysis to evaluate the performance of the five distance metrics. Plots illustrating important points are also included.

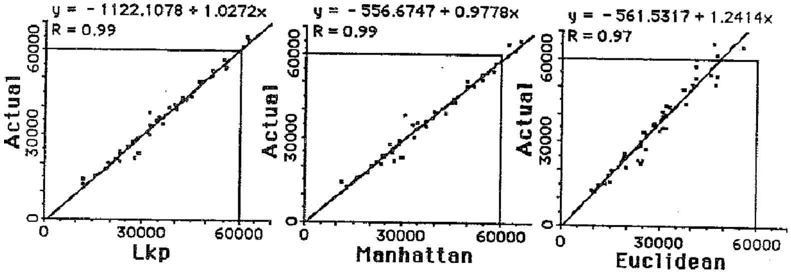
Table 2 : Optimized Parameters and Least Square Root Errors

Area	P	P Error	K	K Error	M Error	E Error	K&P		Error
Estherville	1.08	**2152.7**	1.22	3424.1	2714.2	7689.2	1.06	1.21	**2038.6**
Gulf Port	1.34	12256.1	1.13	12153.1	13466.9	12910.9	1.11	1.83	**12149.5**
Newark	1.21	3838.6	1.18	2911.1	4873.9	5744.7	1.20	2.22	**2888.9**
Salton Sea	1.15	**5092.4**	1.18	5553.8	5441.6	6643.4	0.94	1.03	**5075.1**
Vail	0.86	11012.6	1.44	9062.0	12405.4	21709.5	1.37	1.61	**8852.4**

4.1 Estherville, IA, and Salton Sea, CA

The Estherville quadrangle in rural Western Iowa and the Salton Sea quadrangle in rural Southern California have low local relief and are dominated by agricultural land use laid out in a grid. The Estherville network visually appears more rectilinear than does Salton Sea's. In each case, however, they are not dense, urban grids, but are less dense rural networks that are regular and rectilinear nevertheless. It was expected that the Manhattan metric should out-perform the Euclidean metric in these study areas. As illustrated in Table 2, the RMS value for the Euclidean metric are greater than the Manhattan metric value. The Manhattan metric out-performs the Euclidean metric by a wider margin in the Estherville area, which is consistent with its rectilinear appearance. Also note from Figure 1 that the Euclidean metric consistently provides underestimates when compared to the normative measure.

Figure 1 : Scatterplots for Estherville



The *lp* metric performs better than the Manhattan metric. In the Estherville area, the "p" parameter was optimized at 1.09, the closest to 1.00 (Manhattan) of all "p" values in this study. Since the network is not exactly a rectilinear grid, the inflation of "p" from 1.00 to 1.09 yields the observed greater accuracy. Also in Estherville, the l_k metric yields higher RMS values than both the lp and Manhattan metrics. This is especially interesting in that a non-parameterized metric performed better than a parameterized one. In Salton Sea, the l_k metric has lower error than the Manhattan metric, but is poorer than the lp metric. These results lend support for the argument that the exponential parameter "p" is more sensitive to rectilinearity than is the multiplicative parameter "k." In both study areas the $l_{k,p}$ metric performed slightly better than the lp metric, but not by a significant margin.

4.2 Newark, NJ

The Newark quadrangle in New Jersey exhibits a dense, urban network. Despite this, the Manhattan metric did not perform as well compared with the parameterized metrics as in the rural rectilinear networks discussed above. There were enough diagonal roads, many of them interstates or major highways, to cut actual travel distances below those predicted by the Manhattan model. The distances are still, however, greater than Euclidean. Both the "k" and "p" metrics have lower RMS values than the Manhattan metric, and are optimized at a level midway between the Euclidean and Manhattan levels. The l_k metric is better than the lp metric, but both are surpassed by the $l_{k,p}$ metric as measured by the RMS values.

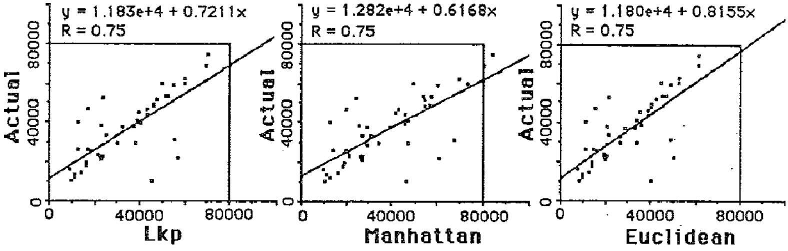
4.3 Vail, CO

The Vail, CO, quadrangle is a mountainous area with few roads, and thus, it is difficult to travel between some places directly. None of the metrics estimate interpoint distances accurately because some actual distances are similar to Euclidean, while others are much longer than Manhattan. Due to this inconsistency in the network, values for "k" or "p" are quite high, as may be seen in Table 2; also note that the $l_{k,p}$ metric yields error half that of the Euclidean. While it is true that the measures we use do not control for total distance, each quadrangle in the study is the same size and has similar distance distributions. The values are much higher in Vail than the three study areas discussed previously. The advantage that parametric estimators have over the Euclidean and Manhattan metrics is their ability to be "fit" to the characteristics of individual networks. When these characteristics vary greatly within a network, this advantage is reduced.

4.4 Gulf Port, MS

The Gulf Port quadrangle in Mississippi and Louisiana also does not clearly support the use of any single metric. This area is located along a coastline, which causes circuitous travel patterns when traveling between some places. The RMS values are high for all five estimators, illustrating their inability to provide accurate distance estimates. The l_k, l_p and $l_{k,p}$ metrics have the second highest RMS values of all the study areas, second only to Vail. Figure 2 illustrates how poorly the functions estimate road travel distances. The regression lines fitted to all plots have slopes which are significantly less than one, illustrating that they systematically under-estimate the travel distances. The r^2 values are less than 0.6 for all functions, meaning that the regression equations explain less than sixty percent of the observed variation in the estimates.

Figure 2 : Scatterplots for Gulf Port



5.0 DIGITIZED DISTANCES

Three lines which vary in sinuosity, shown reduced in Figure 3, were used in a preliminary assessment of distance estimates derived from digitized road networks. The lengths of the lines vary because of differences in sinuosity. The distance estimates in Table 3 were obtained by taking the mean distance values for 10 volunteer subjects. Although the sample of lines is small, it shows that the standard deviation monotonically increases as the sinuosity of the sampled lines increases. This trend is also evident when contolling for the mean line length, by using the coefficient of variation (standard deviation / mean). Additional testing will be required, however, before we can make definitive statements about distance estimates derived from digitized cartographic chains.

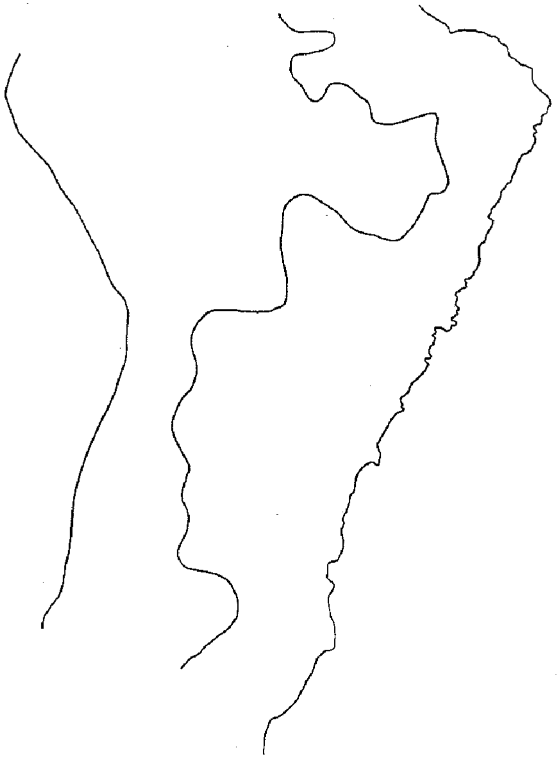
Table 3: Descriptive Statistics for Digitized Distances

	Mean	Std Deviation	CV
High	8152.2	242.8	0.0297
Medium	9790.5	196.9	0.0201
Low	5049.3	70.9	0.0141

6.0 CONCLUSIONS

The distance estimators performed with varying levels of accuracy. Each estimator performed best in homogeneous networks. In irregular networks, all functions exhibit high RMS values. When measuring distances across a network made irregular by natural or other barriers, some alternative should be found. One alternative is to employ a road network model in a GIS (e.g. DLG) to calculate shortest paths.

Figure 3 : Digitized Lines



The $l_{k,p}$ metric has the lowest RMS values in each of our study quadrangles. These findings support Love and Morris' (1972, 1979) contention that parametric distance estimators perform better than the Euclidean or Manhattan metrics. Our results can also lend support the for use of the l_k metric, as argued by Berens and Korling (1985) on the grounds of computational efficiency, since the difference in error between the $l_{k,p}$ and l_k metrics is small. It may be used only, however, in areas for which the network is not rectangular, since the "k" parameter is not as sensitive to rectilinearity as is the "p" parameter. However, we argue that the optimization of two parameters is not technically difficult. In light of this, the use of the $l_{k,p}$ metric is supported.

In the future, GIS technology will play a larger role in spatial analysis and spatial decisionmaking. As the technology to calculate road travel distance improves through increased use of GIS and existing digital cartographic databases, the need to estimate road network distances from points will decrease. Such approaches are not free of error, however, and errors in the context of GIS use must continue to be examined. The preliminary results obtained in the digitizing experiment suggest that variance in distance estimates obtained from digitized road segments increases as a function of line (road) sinuosity. How this effect plays out in a larger sample of roads is a subject that should be examined in further research on distance estimation.

REFERENCES

Armstrong, M.P., Densham, P.J., and Rushton, G. 1986. Architecture for a microcomputer based spatial decision support system. *Proceedings, Second International Symposium on Spatial Data Handling*, Williamsville, NY: IGU Commission on Geographical Data Sensing and Processing. pp. 120-131.

Armstrong, M. P. 1988. Distance imprecision and error in spatial decision support systems. In *Proceedings, International Geographic Information Systems Symposium, Vol 2,* Washington, DC: U.S. Government Printing Office. pp. 23-34.

Berens, W. and Korling, F.-J. 1985. Estimating road distances by mathematical functions. *European Journal of Operational Research*, 21, pp. 54-56.

Davis, J.R. and Grant, I.W. 1987. ADAPT: a knowledge-based decision support system for producing zoning schemes. *Environment and Planning B,* 14, pp. 53-66.

Fildes, R. A. and Westwood, J. B. 1978. The development of linear distance functions for distribution analysis. *Journal of the Operational Research Society*, 29, (6), pp. 585-592.

Francis, R. L., McGinnis, L. F. and White, J. A. 1983. Locational Analysis. *European Journal of Operational Research*, 12, pp. 220-252.

Krarup, J. and Pruzan, P. M. 1980. The impact of distance on location problems. *European Journal of Operational Research*, 4, (4), pp. 256-269.

Kuiper, H. 1986. *Distribution of Distances in Pregeographical Space.* Brookfield, VT : Gower Publishing Company.

Laporte, G., Nobert, Y. and Desrochers, M. 1985. Optimal routing under capacity and distance restrictions. *Operations Research*, 33, (5), pp. 1050-1073.

Love, R. F. and Morris, J. G. 1979. Mathematical models of road travel distances. *Management Science*, 25, (2), pp. 130-139.

Love, R. F. and Morris, J. G. 1972. Modelling inter-city road distances by mathematical functions. *Operational Research Quarterly*, 23, pp. 61-71.

Love, R. F., Morris, J. G. and Wesolowsky, G. O. 1988. Mathematical Models of Travel Distances. Ch.10 in *Facilities Location : Models and Methods.* New York : North Holland.

Lupien, A. E., Moreland, W. H. and Dangermond, J. 1987. Network analysis in geographic information systems. *Photogrammetric Engineering and Remote Sensing*, 53, pp. 1417-1422.

Nordbeck, S. 1963. Computing distances in road nets. *Papers of the Regional Science Association*, 12, pp. 207-220.