Part 4: Mathematical, Algorithmic and Data Structure Issues

GENERALISATION AND ERROR IN SPATIAL DATA BASES

MICHAEL BLAKEMORE, *Department of Geography, University of Durham, Durham, United Kingdom*

THE 115 polygonal units shown in Figure 1 form the 100 Employment Office Areas (EOAs) in North West England used as a geographic base for industrial establishments by the North West Industry Research Unit at the University of Manchester. They are a typical set of units used to partition the continuum of space into more manageable structures, essentially for administrative purposes. The units vary considerably in topographic size, from the very small EOA of Walton (89) to Lancaster (107), differences reflecting not their suitability for spatial analysis but simply the result of administrative convenience whose ideal is that such units should not have too wide a variation in the totals of establishments or workforces within them. The boundaries are merely a result of processes of administrative containment. Their statistical disadvantages have been identified in the contexts of spatial autocorrelation, the modifiable areal-unit problems, or by reorganisation of the data according to impartial units such as grid cells. The latter, though laudable, requires that data be available at a sufficiently fine level for restructuring, and assumes that confidentiality constraints permit such an operation. Usually, however, spatial data are gathered by governmental agencies at local regional and national levels, and these agencies view geographic space as being partitioned not into statistically impartial units but into the administrative structures they know and love. Whatever the relative merits and disadvantages of irregular units they do provide information in a format primarily useful for administrators, and spatial analysts must use them to best advantage.

If the problems only were statistical then those building geographic information systems would have relatively little to worry about. However, there are potentially serious error problems, readily identified in the literature, but few of which have been tested empirically for their total effects in automated spatial retrieval. This is somewhat surprising considering the claims that computer cartography is a highly refined and accurate science. Bickmore (1982) notes various attempts towards scientific processes where any process within it involves a mathematical formulation and algorithmic expression. Dudycha (1981, p. 116) examined 'the computer revolution in cartography' and Morrison (1980,

p. 7) even claims that 'cartographic products produced by computer-assisted technology can usually be made accurate to the resolution of the machine hardware used to produce them.' Various authors in Rhind and Adams (1982) support such views, yet they conflict markedly with authorities such as Jenks (1981), or Boyle (1982, p. 3) who argues that 'in my opinion we have failed during the 1970s.' A major reason must be that of inadequate investigation into error processes, a fault identified by Goodchild (1977) and elaborated by Poiker (1982, p. 241) who notes 'the absence of any notion of precision and accuracy' in computer cartography, which he argues 'is like a person with the body of an athlete in his prime time and the mind of a child'; Goodchild (1980a, p. 192) examined accuracy for raster data, stating

> The accuracy problem would be simple if measurements for digital data could be checked directly against their true, real world values. But this is not normally possible. In general, accuracy must be predicted from the digital data alone, by making assumptions about the true data.

Such assumptions relate to data usually 'captured' from published paper maps (with all their imperfections), using highly sophisticated hardware.

Poiker's 'body of an athlete' is largely undisputed. Without doubt the hardware of computer cartography is becoming faster, cheaper, and more accurate in a mathematical sense. What seems less easily agreed is the 'mind of a child,' which here is examined in the context of misuse of error-prone cartographic data. The sizes of the EOAs in Figure 1 do not reflect the numbers of establishments within them – not surprising and Tobler's (1979) suggestion of a 'transformational view of cartography' should motivate more to think in other metrics than raw topographic domains. In the context of EOAs the topographic size is more indicative of dispersion – the larger the EOA the more dispersed the establishments – and this has been confounded by the ravages of time and industrial recession to give wide variation in establishments and employment totals. EOA 107 (Lancaster) had 209 recorded manufacturing establishments in the database with an average employment of 34, whereas the much smaller EOA of Manchester (57) had 2367 establishments, averaging an employment of 11. The topographic boundaries pay scant regard to the thematic variables, yet it so often is the case that the topographic dimension is the primary mode of geocoding. A standard approach uses an orthogonal grid framework to geocode boundaries (usually as segments/chains, nodes, features, points) and utilises retrieval algorithms such as point-in-polygon to extract requisite locations. Considerable research has gone into optimising the search time of such algorithms, by refining software codes and structuring databases to minimise disk accesses. Both approaches seek to utilise the power and numerical accuracy of the computer more efficiently. The eventual cartographic precision, however, is determined by the quality of input data and types of usage. Jenks (1981) highlights various error inputs of digitising, and Chrisman (1982) assesses the error components of maps, though their studies are relatively recent. Early work on retrieval by point-in-polygon was concerned mainly with execution speed (for example Aldred 1972, Nordbeck and Rystedt, 1972). Baxter (1976) seems to assume

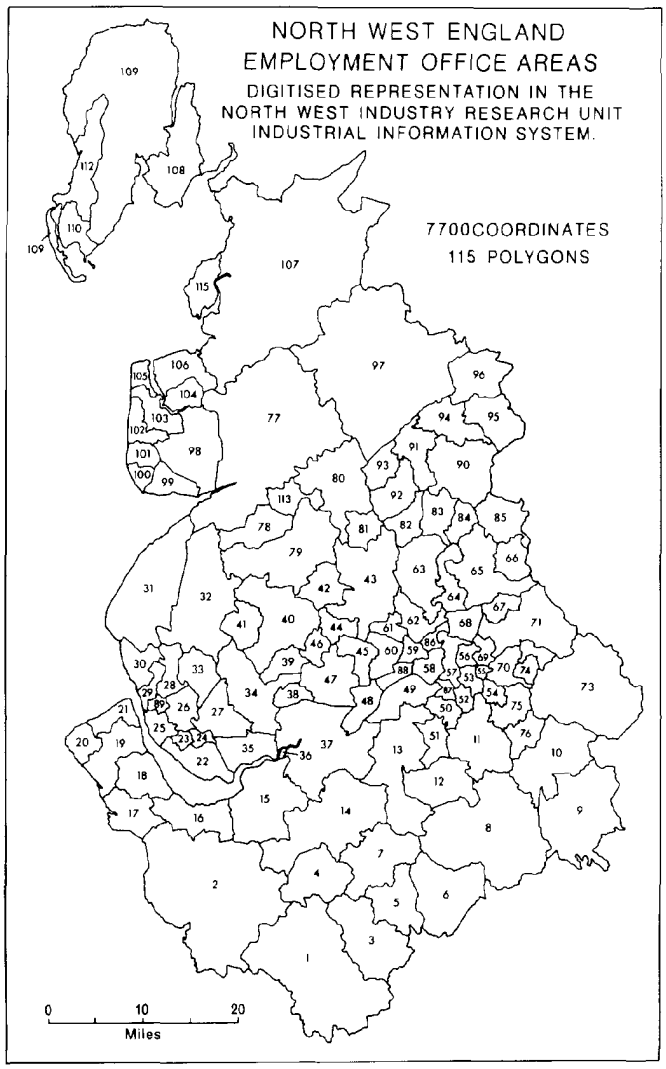

FIGURE 1. *The 115 polygonal units shown form the 100 Employment Office Areas in N.W. England.*

inherent geocoding precision when stating that for any retrieval process 'the user merely states the precise coordinates of ...' the features in question. There can be no precision in map-derived data particularly since so many boundary features used are artificial lines that cannot be verified for ground truth. Precision also is underplayed by the Department of the Environment's (1973) extensive document on geocoding, which states as a matter of fact that 12-figure Ordnance Survey coordinates will give a ground resolution of 1 metre on the ground and

that this will be 'adequate for most Local Authority data processing purposes.' Furthermore, they claim

> It is normally possible to establish point-coordinates by eye using 1 : 1250 maps within an accuracy of around plus or minus 5 metres. When digitising equipment is used to generate coordinates electronically, the accuracy is improved to within plus or minus 1.5 metres (DoE 1973, p. 93).

In spite of the fact that the same human eye is guiding the digitiser cursor! Given a 1mm line on a 1:1250 map, a paper map with potential maximum stretch of 5%, the line would have a maximum distortion of 0.0625 metres on the ground. The 1mm line itself represents a direct width of 1.25 metres and the potential ground truth is 1.3125 metres. The 'human frailties' (Jenks, 1981) of digitising would add further error to this (for a statistical evaluation of error distributions see Chrisman, 1982) and there is little chance that the 1 metre resolution would be achievable. It seems logical to regard all digitised lines as being error-prone. Finer digitising would involve a futile attempt to transmit computational precision of the machine to the vagaries of cartographic representation.

There is also another, and very variable error component related to geocoding and digitising. Aldred (1972, p. 5) regarding digital polygon representation, notes 'the accuracy of this representation for curved shapes being dependent upon the number of vertices used.' Goodchild (1980a) quotes the paradoxical situation whereby greater generalisation in digitising gives less accuracy but the smaller volume of data allows faster processing. He also argues (Goodchild 1980b, p. 89) that cartographic generalisation may vary from map to map on the same scale. Even on a single map the digital sampling error will be distributed unequally – lines that are straight will have relatively low error components, while crenulate lines will be seriously error-prone. These differences are worsened by effects of unit size, particularly since the smaller sized spatial units in this study contain disproportionately high numbers of establishments. Thus local error factors will exacerbate global errors.

Smedley and Aldred (1980) examined these error sources. They note that the translation of a continuous line on a map into a digital summary involves a radical change in dimensionality. Of the infinity of possible points along any line, digitising samples but a few. Shapes become simplified, lines have their paths generalised, and to confound all of these there are the human problems in digitising. In spite of a growing range of hardware at the upper end of the market – line-followers, scanners etc., – much digitising still is undertaken using conventional tables, and the majority of existing line files are so derived. In Universities, Local Authorities and Research Organisations, digitising has been the poor relation in geographic information processing. Reasons are not difficult to isolate, since it is a tedious, time-consuming and exhausting process with a low reward value. It is easy to impress with a multi-coloured computer map, much less so with a slide of clean digitised outlines. It has been a case of out-of-sight out-of-mind and this mentality has been one reason for insufficient appreciation of problems of resolution and reliability in retrieval situations.

One technique for assessing geocoding and retrieval error is the 'epsilon'

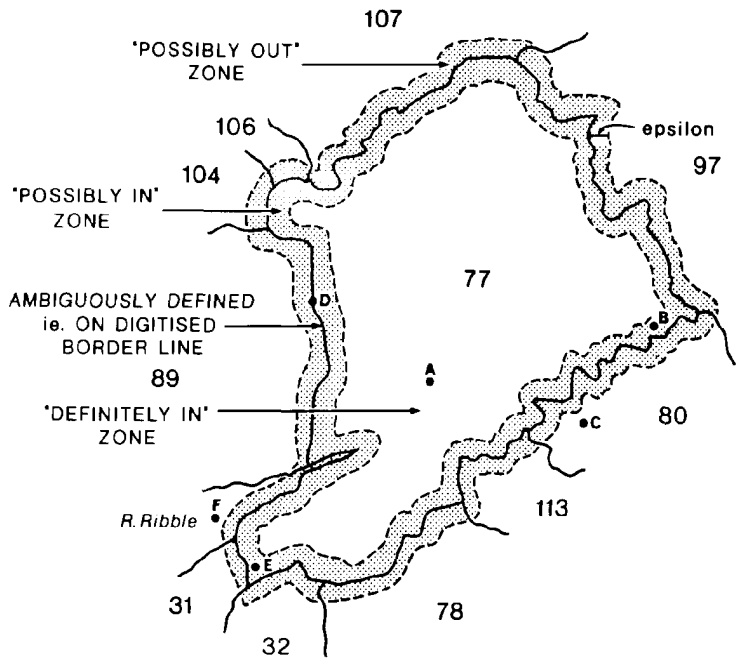

FIGURE 2. *Epsilon bounded point in polygon check.*

distance of Perkal (1966). Perkal used a band of error (distance epsilon) about a cartographic line as a means of generalising the line objectively. The technique works as efficiently in reverse if, for example, one simple value of epsilon is seen in the context of the representative fraction of a map. If the scale is 1:50,000 and the width of a line 1.2mm then the line cannot be measured to a ground precision of better than 60 metres. Figure 2 shows an epsilon error band placed above EOA 77 (Preston). The black lines are the digital representation and because of the various error processes a band of error is assumed to exist.

The area contained within the confines of the band now has four sectors. 'Definitely in' is the core of the area within the error tolerance. 'Possibly in' records a point which is within the digitised confines of EOA 77 but the error band makes inclusion uncertain. Beyond the digitised line is 'possibly out' where points are geocoded as being in a neighbouring area but technically they could be EOA 77's members erroneously geocoded, mis-classed because of digitising error, or both. Lastly, an establishment may be geocoded actually on a digitised line. The chances of this may seem remote, but they do occur and few algorithms allow for this – Douglas' (undated) being one of those that does. Thus on Figure 2, point A will be exclusively assigned to EOA 77, and point C to EOA 80. Point B is possibly in 77 and therefore possibly out of 80. Point D is ambiguously defined being on the border of 89 and 77. Point F is a more complex situation and will be discussed later.

An epsilon bounded point-in-polygon check can be carried out easily using a

combined point-in-polygon and point-in-path check, where the width of the path is twice the value of epsilon. It is important that a point-in-polygon algorithm be used which specifically checks for ambiguity. The value of epsilon will be application dependent (see Chrisman 1982, for statistical arguments), but if it highlights a large percentage of doubtful assignments then a different approach to automated spatial retrieval would be necessary.

To evaluate the effects of the error process an initial test was carried out in 1982 and 780 initial entries in the NWIRU database. These industrial establishments provide a readily identifiable epsilon of 0.7071km because they are geocoded by the Department of Industry to a 1 kilometre grid square resolution. This may sound somewhat coarse, but some manufacturing units can have sites in excess of one square kilometre. Additionally, it is standard practice to assign a single coordinate pair to spatial features of areal extent because coordinates can be stored easily with other data, and they minimise retrieval time. Point geocoding is simple and direct, but it is not used for any methodologically sound spatial reason, simply because the characteristics of digital computing dictate it. Now that extremely powerful processors are becoming more common, it may be feasible to propose a reconsideration of geocoding practice. After all, current methods are the results of compromises and constraints determined by technology that is now decidedly out-of-date. Current practices are throwbacks to earlier days, although given the large amount of investment in these practices a reconsideration would not be entirely appreciated.

The 780 test points were not chosen with any particular areas in mind, although they did cover an area of eastern Manchester with EOAs of varying sizes; EOAs 73, 76, 11, 53, 57, 58, 68, 67 and 71 bounded the zone. EOA 73 is a large, almost circular area whereas 55 is small, and 70 has a promontory which would be almost all within the epsilon band. All the 780 points were tested against every EOA for various levels of epsilon so that sudden declines in accuracy could be identified. Starting with a classic point-in-polygon search, 23 of the 780 were flagged as ambiguous, 2.95% of the total. With an epsilon of 0.1km 7% of all points were in areas of doubt: 100 metres is less than the ground-thickness of a line on most administrative maps and, given statistical tolerances of 5%, indicates that careful checking for erroneous geocoding is increasingly important. At 0.2km 20% were doubtful, a large increase, and for epsilons of 0.3km to 1km the percentages in doubt were 23, 27, 35, 44, 50, 52, 56, and 61.

At an epsilon of 0.7km only some 50 percent of the 780 locations are uniquely assignable to one area. A worrying aspect is that the percentage loss is not equally distributed between the EOAs; 71 and 73, which are large and which do not have many inflections, are least affected. At an epsilon of 1km they had respective losses of 32 per cent and 33 per cent. In the final NWIRU database, these were to have 709 and 17 establishments respectively. Compare this situation with areas 55 and 56 which suffered 100 per cent doubt at epsilons of 0.5 and 0.4 km respectively. These smaller units, which, as mentioned previously, tend to contain more than their 'topographically fair' proportion of establishments, are most at risk from cartographic error.

Overall, the initial trial highlighted a need for careful appraisal of cartographic

error in digital spatial retrieval. First, the areas which are most affected by epsilon are the smaller areas, areas which are often those units of most importance in a geographic information system. They are the urban areas wherein most activity is concentrated. Second, shape has an effect of its own which can compound the effect of unit size. A small unit which has considerable elongation would totally fall within the doubtful areas of the error band. Third, the rate of error may also be a function of establishments locating near administrative boundaries.

The test area was examined further to see what the effect of epsilon would be on employment totals, not just on numbers of establishments. An alarming range of values occurred for average employment per establishment using only 'definitely in' establishments. Interestingly, the deviations were not necessarily highest in the smallest-sized units. This simply is because of the variation in size of employment of establishments, and it only requires one large employer to be in the zone of doubt and the averages for the particular areas affected will vary markedly. One possible counter-argument to this would be to put faith in the presence of spatial autocorrelation. This would view the epsilon error very much as a case of swings and roundabouts, whereby the establishments that EOA 11 'loses' to EOA 8 are offset by those which EOA 11 'gains' from EOA 8. Therefore, a further simulation was carried out using 1980 total employment and, instead of excluding the doubtful establishments the possibly in and possibly out establishments were included. Only EOA 73 remained unchanged at an average employment of 33. The smaller areas, however, suffered badly: EOA 53 had averages ranging from 4 to 14.91 and 56 from 9.7 to 15.5.

Here the error problem was attacked by 'hedging one's bets' and rather than ignoring establishments in the error band, including all of them on either side of the border. There is a greater chance, perhaps, of ironing out some of the error by a neighbourhood factor, but it could be argued that this is no more than a cynical spatial lag whereby some of the establishments in the areas nearest neighbours are being used in an attempt to reduce the variability induced by epsilon error. In this case the greater the value of epsilon the greater the smoothing, and ironically, therefore, the bigger the error term the less the error will be visible in the smoothed data. Further, some EOAS started to 'gain' employment at an alarming rate and the results did not warrant any further consideration.

The outcome of the initial testing on the 780 points was that a decision was made to validate the entire database. The possibility of error was far too high to be ignored. The results of the entire epsilon testing of the database at 0.7071km are listed in Table 1.

The categories bear some explanation. 'Possibly out definite' refers to those establishments which, using conventional point-in-polygon retrieval, would be missed altogether. This sort of case occurs along coastal areas where, because of the establishment geocoding to 1km resolution, the grid intersection lies out in the sea. Only the 'possibly out' epsilon can pick this up. However, since there is no other EOA to which the establishment can be assigned, it can be uniquely assigned to the EOA for which it is 'possibly out.' Category 2 is similar – the

Table 1 FULL DATABASE (22,798 ESTABLISHMENTS) EPSILON ERROR RESULTS AT 0.7071 KM

Category		Percentage Affected
1. Possibly out definite		1.5
2. Possibly in		4.35
3. Unassignable		1.4
4. Possibly in/out		29.8
5. Possibly in/out 1		6.72
6. Ambiguous		1.19
	Subtotal	44.96
7. Uniquely assigned		55.04
	Total	100.00

establishment is possibly in an EOA but there are no other EOAs for it to be 'possibly out.' Referring to Figure 1, this could involve establishments at the margins of EOAs 8, 9, 73, 71, and so on. Again, they are uniquely assigned. Category 3 is unassignable. On Figure 2, location F is one such example, since it is beyond the error tolerates of all surrounding EOAs. Category 4 refers to establishments which are flagged as possibly in one and possibly out of another EOA. Category 5 notes those which are possibly in one and possibly out of more than one other EOA and 6 refers to those establishments by chance geocoded on the digitised boundary. The establishments flagged in Categories 3 to 6 were all checked against detailed records and local street maps in a time-consuming but important check as to their correct EOA. In a large number of cases this check indicated that traditional point-in-polygon retrieval would have been manifestly unreliable and would have produced classifications of establishments which at times bore little resemblance to the truth. At e = 0.7071km the 'definitely in' category only will include some 60 percent of establishments and workforce. It is useful to note that the number of establishments closely follows the trend of workforce and that the numbers of establishments can be used as a reliable surrogate for other key variables.

The end result of the NWIRU's concern with map error and spatial retrieval was a considerable amount of checking and manual validation but a database which verified as 100 per cent accurate at the level of Employment Office Areas. For any aggregation or combination of EOAs this consistency can be maintained. Since every establishment is uniquely assigned, it seemed useful to include the EOA assignment as an extra item of data for each record. This extra item of data then was converted into a link list format so that each establishment points to the next establishment in the same EOA, so facilitating very high speed retrieval of information without further recourse to point-in-polygon and all the error that it entails. Clearly, not every spatial search will be along the lines of a neat aggregation of EOAs and it must be accepted that irregular area retrieval will be necessary at some stage. For the moment, users can be given three statistics for each such search – these relate to 'uniquely in', all establishments within the digitised boundary, and the total using the autocorrelation effect. Using existing pro-

gramming styles and retrieval techniques, this seems, at present, to be the most logical provision. Nevertheless it does point to important, and indeed urgent, future research into *Intelligent* Geographic Information Systems: systems which can be provided with basic ground roles of spatial inclusion/exclusion that go beyond the crude mechanical techniques in use today. Already Image Analysis researchers are examining developments in 'Context Analysis' which will help to classify satellite imagery using statistical classification techniques, tempered by behavioural inputs from human operators; behavioural inputs which the system learns and will implement automatically at a later date. Such developments will be needed in Spatial Information processing before Geographic Information Systems can operate with the subtlety of a researcher rather than being a brute force speeding-up of repetitive and tedious operations.

REFERENCES

ALDRED, B.K. 1972. *Point in polygon algorithms*, Peterlee, U.K., IBM LTD.

BAXTER, R.S. 1976. *Computer and statistical techniques for planners*, London, Methuen.

BICKMORE, D.P. ed. 1982. *Perspectives in the alternative cartography: cartographic computing technology and its applications* (Cartographica 19, 2).

BOYLE, A.R. 1982. The last ten years of automated cartography: a personal view, pp 1–3 in Rhind and Adams.

CHRISMAN, N.R. 1982. *Methods of spatial analysis based on errors in categorical maps.* Unpublished PhD thesis, University of Bristol.

DEPARTMENT OF THE ENVIRONMENT 1973. *Manual on point referencing properties and parcels of land*, London, HMSO.

DOUGLAS, D. undated. *Collected algorithms*, Cambridge Mass., Lab for Computer Graphics and Spatial Analysis.

DUDYCHA, D.J. 1981. The impact of computer cartography. *Cartographica* 18, 2, pp 117–150.

GOODCHILD, M.F. 1977. Statistical aspects of the polygon overlay problem: in Dutton, G. ed, *Harvard Papers on Geographic Information Systems: vol 6*, Reading, Mass., Addison Wesley.

——— 1980a. The effects of generalisation in geographical data encoding: pp 191–205 in Freeman, H. and Pieroni, G.G. eds *Map Data Processing*, N. York, Academic Press.

——— 1980b. Fractals and the accuracy of geographical measures: *Mathematical Geology* 12, 2, pp 85–98.

JENKS, G.F. 1981. Lines, computers and human frailties, *Annals A.A.G.* 71, 1, pp 1–10.

MORRISON, J.L. 1980. Computer technology and cartographic change: pp 5–23 in Taylor, D.R.F. ed, *The computer in contemporary cartography*, Chichester, Wiley.

NORDBECK, S. & RYSTEDT, B. 1972. *Computer cartography. The mapping system* NORMAP, Lund, Studentlitteratur.

PERKAL, J. 1966. On the length of empirical curves. *Discussion Paper 10 Ann Arbor* MI, Michigan Inter-University Community of Mathematical Geographers.

POIKER, T.K. 1982. Looking at computer cartography, *GeoJournal* 6, 3, pp 241–249.

RHIND, D. & ADAMS, T. eds. 1982. *Computers in cartography*, London, British Cartographic Society.

SMEDLEY, B. & ALDRED, B. 1980. Problems with geodata. pp 539–554 in Blaser, A. ed, *Data base techniques for pictorial applications*, Berlin & New York, Springer-Verlag.

TOBLER, W.R. 1979. A transformational view of cartography. *American Cartographer* 6, pp 101–106.