Assessment of error in digital vector data using fractal geometry

Matt Duckham and Jane Drummond
Department of Geography and Topographic Science
University of Glasgow
Glasgow
G12 8QQ
UK

13 January 1998

Keywords

error assessment; fractal geometry; map scale; internal evidence;

Short title

Error assessment using fractal geometry

Acknowledgements

We are grateful to the Natural Environmental Research Council (NERC) for funding; MJD is currently supported by a NERC CASE studentship on error in GIS, the CASE partner being Survey and Development Services (SDS), Bo'ness, UK. Ordnance Survey of Great Britain (OSGB) data has been supplied under the CHEST scheme. In addition the authors would like to express their appreciation to Stewart McNeill at the Department of Computing Science, University of Glasgow, for technical support. Finally, the comprehensive and constructive comments of all the reviewers were instrumental and invaluable in shaping this paper.

Correspondence address

Matt Duckham
Department of Geography and Topographic Science
University of Glasgow
Glasgow
G12 8QQ, UK

Email: mduckham@geog.gla.ac.uk
Tel: 0141 330 4782
Fax: 0141 330 4894

Assessment of error in digital vector data using fractal geometry

Abstract

This paper presents a new method for assessment of error in digital vector geographic data, where the features represented can be modelled closely by fractal geometry. Using example hydrological data from Ordnance Survey of Great Britain maps at a range of scales, a resolution smaller than which the digital representation of the feature does not exhibit fractal characteristics can be calculated. It is proposed that this resolution reflects the minimum ground resolution of the map, which in turn can be related to the source map scale.

1 Introduction

The inclusion of error management facilities within Geographic Information Systems (GIS) will have a profound effect upon the future industrial application of and research into GIS. Error is described by Chrisman (1991) as a 'fundamental dimension of data', yet most commercial systems provide no error management. In recent years, considerable research effort has been directed towards the assessment of error (Brunsdon and Openshaw 1993; Brunsdon and Openshaw 1994), error propagation (Openshaw, Charlton, and Carver 1991; Veregin 1995), and visualisation of error (van Elzakker, Ramlal, and Drummond 1992; Brown and van Elzakker 1993; Howard and MacEachran 1996; Spear, Hall, and Wandsworth 1996). This research is already leading to some limited tools being supplied with off-the-shelf GIS (Drummond, Brown, Du, and van Elzakker 1996), but as research continues the basis for what has been described as an *error sensitive* GIS (Unwin 1995) should be extended.

Error can be defined as the difference between an observed or calculated value and the 'true' value. The reference to 'true' used in this definition, however, requires further clarification. Any observation of the physical world outside occurs within the context of some model or abstraction (Raper and Livingston 1995): 'theory underlies the taking of any observation in science' (Haines-Young and Petch 1986). The theory, model or abstraction used to describe the 'true' or ideal data set is often termed the *terrain nominale* (Aalders 1996; David, van den Herrewegen, and Salgé 1996). It is the discrepancy between the 'true' values described by this terrain nominale and a particular data set which constitute the error we usually wish to be able to manage and manipulate. The relationship between the physical world, terrain nominale and a particular data set is illustrated in Figure 1.

Data quality statistics are the usual mechanism for quantifying error. Any error sensitive GIS should be able to store and manipulate error in the form of data quality statistics and be able to display and communicate to the user the error represented by the quality statistics. Unfortunately, in most cases such quality information does not exist. Much of the research into assessment of error, and so the production of these quality statistics, focuses upon the National Committee for Digital Cartographic Data Standards' (NCDCDS) preferred method of comparison with an independent source of higher accuracy (NCDCDS, 1988). Here the sources of higher accuracy are taken to be 'true' and to form the terrain nominale. Such sources of higher accuracy are usually unavailable and most data users would not expect to be in a position to collect these data sources themselves. The increasing popularity of sub-decimeter accuracy GPS procedures and of user-friendly GPS receivers and software amongst land surveyors may help to erode this information gap, allowing independent sources of higher accuracy to be collected by data users more frequently in the future (Pohl 1996; Stanislawski, Dewitt, and Surestha 1996). However, it is still desirable to look to other methods of producing data quality statistics, since such methods can act as a primary source of quality information where independent data sources are unavailable or as support for quality statistics where they already exist.

A limited number of methods have been formulated to assess quality without the need for comparison with sources of higher accuracy. Instead these methods rely on the formulation of a suitable mathematical model of the terrain nominale, which is then compared to the actual data to assess data quality. These methods are evaluated in the next section under the heading of *internal evidence*, whilst the main focus of this paper is to introduce a new internal evidence methodology which makes use of the fractal model as a terrain nominale to assess the quality of digital vector data.

2 Background

2.1 Internal evidence

Internal evidence supports the process of assessing and quantifying errors, based on the analysis of discrepancies between a specific data set and a model describing a suitable generalised or ideal dataset. This model is the terrain nominale. Quality assessment through internal evidence, then, operates by formulating a mathematical model which describes the terrain nominale, comparing the data set to this terrain nominale and making inferences about the quality of the data set based on any inconsistencies. The advantage of using internal evidence is clearly that the data itself contains the evidence, so no further data collection for assessing data quality is required.

A considerable and well established body of work already exists in the use of internal evidence, in the form of least squares adjustment. Discrepancies between geomatics observations, including observations from survey and photogrammetry, and the ideal dataset are utilised in least squares adjustment. Some less well established, novel methods for determining data quality using internal evidence have also appeared in recent years. Both established and novel methods are explored in the following sections.

2.2 Geomatics and adjustment

When control points are surveyed to provide the framework for subsequent mapping, they are usually adjusted using least squares techniques to provide a best estimate of their x, y and z values (easting, northing and height) (Bannister, Raymond, and Baker 1992). In addition to providing the best estimate of the control points' x, y and z coordinates, least squares adjustment also provides an estimated precision of the adjusted coordinates in terms of a standard deviation (Mikhail 1978).

Having established best estimates of control point coordinates and their precision, the mapping task proceeds. Formerly, this resulted in hardcopy products which might subsequently have been digitised, but more recently in the direct digital storage of the coordinates of map detail. In either case, the error in the original control point coordinates and that introduced by subsequent mapping of detail will be propagated by the processes the data undergo, for example coordinate transformation. This propagation can be modelled, using variance propagation techniques, to estimate the precision of every mapped (Drummond 1995).

Within the GIS environment the coordinates of terrain features may be used in such information generation tasks as area or perimeter computation. Again, variance propagation techniques can be applied to estimate the precision of the results of such computations, as long as either the precision of each of the stored coordinates forms part of the meta-data of the geospatial database or there are sufficient quality statistics and lineage details that full quality statistics can be generated on demand.

Currently, neither the necessary meta-data nor adequate lineage information is usually found in geospatial databases. Thus estimating the precision of the results of an information generation task using variance propagation is not a viable option *unless* assumptions can be made about the quality of the data. Guidance supporting such assumptions can be found (Harley 1975; Newby 1992) but requires a knowledge of the data source. Identifying the source may be straight forward; a region, such as the English Lake District, will have been mapped by a few well know UK map publishing organisations, eg the Ordnance Survey of Great Britain, the Automobile Association, Harper-Collins, Harveys Ltd. Each of these publishers produces maps from which distinctive digital datasets tend to arise in terms of the number of road classes, railway classes, administration boundaries etc supported. Thus, the data's feature classes or attributes may indicate their source and hence a certain data quality can be assumed.

However, given that much of the popularity of GIS lies in its ability to integrate data, it is to be expected that for integrated data sets data source may not be known in any useful way. This undesirable situation is especially likely when a national mapping organisation is not yet supplying full coverage on a standard reference frame: data may come from anywhere and along tortuous routes. If identifying the source is not straightforward and if appropriate meta-data do not exist, is there other internal evidence indicating source or quality?

2.3 Novel approaches to internal evidence

There are also a limited number of novel models which can be used to assess data quality based on internal evidence. Probably the most elegant method uses the frequency histogram of a digital elevation model (DEM). Wood (1996) found that the frequency histogram for most Ordnance Survey DEMs tested exhibited a peaked distribution, where elevations of multiples of 10m occurred at as much as twice the frequency of elevations between these multiples. In contrast, the frequency distribution of elevations are expected to be smooth. This discrepancy between the expected model and the actual data was attributed to the interpolation process used to produce the DEM from the original contour data and to the inevitable statistical over-representation of particular elevations within contour data. In terms of data quality, the inconsistency between the model and the physical world allows us to make some assumptions about the lineage of the data; that it has been produced by interpolation from contour lines at 10m spacing. Further, the magnitude of the peaks could potentially facilitate assumptions about the accuracy of the data.

Polidori, Chorowicz, and Guillande (1991) have made use of fractal geometry to assess DEM quality. Fractals exhibit self-affinity, the property of shapes which can be divided into sub-sets and linearly mapped onto the whole. This property has been termed scale-independence (Clarke 1986) since self-affinity means the same features appear at all scales (Datcu and Seidel 1994). Fractal dimension is a measure of the degree of self-affinity of a shape. (Polidori, Chorowicz, and Guillande) measured the fractal dimension of a DEM for large and small scales. Because fractals exhibit scale-independence, the fractal model would suggest that, given that topography is fractal, the fractal dimension of the DEM should exhibit no variation with scale. The research uncovered much lower fractal dimensions for large scale data than for small scale data, indicating that the DEM did not exhibit fractal behaviour over small distances. This discrepancy was again attributed to the interpolation process used to create the DEM from contour data, which is responsible for excessive smoothing for distances smaller than the horizontal contour interval. While the method is simple, powerful, and even adaptable to assessment of anisotropy in the DEM, Goodchild and Tate (1992) note its limitations. Primarily, the method depends heavily on the the assumption that topography is fractal. There has been some debate over whether this assumption is reasonable, many researchers conceding that topography is fractal, but that a single fractal dimension only exists at limited scale ranges and over limited areas (Clarke 1986; Goodchild 1988; Polidori 1994). (Polidori, Chorowicz, and Guillande) accept that the atypical or non-fractal nature of topography could provide an alternative explanation for the locally planar characteristics of the DEM, but maintain that the coincidence of the excessive smoothing and the contour interval is suggestive of link a between the two.

Brunsdon and Openshaw (1993, Brunsdon and Openshaw (1994) have used an index based on a Haar transform to classify line segments and produce a local measure of line complexity. By modeling digitising error as dependent on geometric complexity, the method produces a local measure of the expected accuracy of each line segment. The method is of particular interest because it is able to partition a digitised line into individual line segments. The authors maintain that the length of the original digitised lines will be of arbitrary length, whereas the partitioned line segments will be homogeneous with respect to complexity. The technique suffers partly because arguably it is only applicable to digitised line features and because the index of line complexity produced has no absolute value. Hence it can only be related to accuracy through empirical relationships.

Other approaches can be identified. Regionalised variable theory, which decomposes the spatial variation of a variable into a structural component, a spatially correlated error component and a residual, spatially uncorrelated error term (Heuvelink, Burrough, and Stein 1989) can also been suggested as a model which could be used to yield quality through internal evidence. Atkinson (1995) has used a modified variogram to estimate the information content of a continuous field-based data set from which the spatially uncorrelated error term can be deduced. Although in a minority of cases there may no spatially correlated error (Monckton 1994), this analysis is unlikely find significant use in assessing data quality since it is usually the spatially correlated error component that is of greatest interest.

Some procedures are already in use. For example, the distribution of non-spatial attributes within a mapped area can be predicted based on established land use distribution models and the assumption that deviation from the models indicates error. Such procedures are now so well established that they are applied in project planning by the Ordnance Survey of Great Britain (OSGB, 1997). All the methods for assessment of data quality through internal evidence reviewed here compare the data with some model, formulated to describe the data given a number of assumptions about the data. There are, however, only a handful of suitable models available at present, most focusing on continuous field data, such as elevation. The methodology presented in the next section builds upon that put forward by Polidori, Chorowicz, and Guillande (1991) and is based on the comparison of the fractal characteristics of vector hydrological data with the fractal model. The quality information produced by the method, often otherwise unavailable, is of the form required for the execution of least squares adjustment.

3 Application of the fractal model

As emphasised previously, the formulation of a terrain nominale inevitably entails a number of assumptions be made. The successful application of any quality assessment which uses internal evidence will depend primarily on the degree to which these assumptions are reasonable and met for a given data set. The vector data model, used in many GIS, models lines and boundaries as a locally linear curve. Tveite and Langaas (1999) point out that any fractal behaviour for infrastrucure and anthropogenic features is expected to break down at large scales. Similarly, the digital vector representation of natural fractal features will exhibit fractal behaviour only within limited bounds. Given a feature which is well modelled by fractal geometry, it is possible to determine the resolution at which the fractal nature of the curve breaks down, the curve succumbs to the vector data model and becomes linear. This section first details the assumptions made when using a fractal model as the terrain nominale, and then sets out the methodology for determining this resolution.

3.1 Importance of fractalness

'True' fractals exhibit scale-independence; the existence of detail at every scale. It follows that no geographic feature can be perfectly modelled by a fractal since all geographic features are inevitably constrained by upper and lower limits (Pentland 1984). The foremost criticism of work of Polidori, Chorowicz, and Guillande (1991) by Goodchild and Tate (1992) was that topography was not modelled particularly well by conventional fractal models. This criticism cannot be taken too seriously; at the heart of any application of fractals to environmental data is the assumption that the features represented by the data are well modelled by fractal geometry. Xia and Clarke (1997) highlight many of the potential pitfalls of the inappropriate application of the fractal model. The assumption of 'fractalness' is carried through into the methodology presented here; the results of data analysis using this methodology will only be valid as far as the features represented by the data can be modelled by fractal geometry. Having said this, many authors have argued that most environmental data displays scale-independence (Burrough 1981; Mandelbrot 1982; Clarke 1986), certainly it is the case for many environmental features that fractal geometry presents a *better* model than Euclidean geometry.

3.2 Other possible assumptions

Goodchild and Tate (1992) make two further criticisms of the work of Polidori, Chorowicz, and Guillande (1991) which need to be addressed before any further work based on Polidori, Chorowicz, and Guillande (1991) can be undertaken. Firstly, the tendency of the calculation of fractal dimension to vary with the choice of analysis technique needs to be considered when evaluating the results. Whilst this in undoubtedly the case (Peitgen, Jurgens, and Saupe 1992; Xia and Clarke 1997), the analysis presented in the following section (§3.3) consistently uses the same analysis technique and is focussed on determining when this technique fails rather than on the magnitude of the fractal dimension. Secondly, the methodology presented by Polidori, Chorowicz, and Guillande (1991) uses a sub-sample of profiles from the DEM surface to make inferences about the fractal nature of the surface as a whole. Goodchild and Tate (1992) point out that fractal theory offers no guidance on the confidence in or variability of the fractal dimension of such samples and so cannot support the assertion that the results have not arisen by chance. This criticism does not apply to the analysis in the following section, since the analysis can be applied to the entire feature in question and no sub-sampling of the feature is required.

3.3 Measurement of fractal dimension

A variety of methods exist for calculating the fractal dimension of a curve. The simplest are dividers methods, which resolve shape as a function of scale (Carr and Benzer 1991). The length of a fractal curve is dependent on the scale at which it is viewed. The observed length $L(\delta)$ of a fractal curve is related to the number of steps of size δ required to describe the curve by Equation [1] below, where a is a constant of proportionality and D is the fractal dimension of the curve.

$$L(\delta) \quad = \quad a.\delta^{1-D} \qquad \text{(Feder 1989)} \tag{1}$$

However, in the case of digital vector geographic data stored as a series of coordinate pairs this relationship is expected to break down for small values of δ. Despite the assumed fractal nature of the geographic features being represented, the actual curve will have a limited resolution as a consequence of the finite resolution of all measurement and data capture equipment. Consequently, for very small values of δ the observed length of the line will not increase without limit, as suggested by [1], but will approach the Euclidean metric length of the curve [3] and indeed become independent of δ [2].

$$\lim_{\delta \to 0} L(\delta) = d_E(x,y) \tag{2}$$

$$\text{where} \qquad d_E(x,y) = \sqrt{\sum_{i=1}^{N} (x_{i-1} - x_i)^2 + (y_{i-1} - y_i)^2} \tag{3}$$

At some critical value or range of δ, therefore, we would expect the behaviour of the curve to switch from the fractal behaviour described by [1] to the linear behaviour described by [2]. Furthermore, since we have made the provision that the geographic feature being represented by the curve is fractal, this critical value λ will be the resolution of the original data. In practical terms, no information exists about features at a smaller resolution than λ.

3.4 Data sets

The data used in this study was OSGB digital data of the English Lake District. As detailed in §3.1 the assumption that the features analysed are modelled well by fractal geometry is central to the successful application of the analysis. Hydrological features were chosen to be extracted from the data since rivers in particular are very well modelled by fractal geometry, being amongst the classic natural fractals in Mandelbrot's 'Fractal Geometry of Nature' (Mandelbrot 1982). Since Mandelbrot's pioneering work, a number of more detailed studies have investigated the fractal nature of river networks (La Barbera and Rosso 1989; Robert and Roy 1990; Phillips 1993). Whilst the focus of most of this research has been the physical significance of the fractal dimension of the river networks and its relationship to the area of a river network's drainage basin, all the work is supportive of the assumption that river networks are well modelled by fractal geometry.

OSGB data derived from maps at a variety of scales was tested, namely 1:1250, 1:2500, 1:250 000 and 1:625 000 maps. Tiles from national grid reference NY0026NE, NY0427, NYSW and NY were chosen from each scale respectively for reasons of availability and because they contained an appreciable portion of the local river network. From each of the tiles the river features were extracted. Due to the range of scales being studied it was impractical to attempt to use different representations of the same river at all scales. Hydrological data at the four scales does not use the same classification, this study using 'Water Feature' (feature code 0059) for 1:1250 and 1:2500 scale, 'Minor River' (feature code 5230) for 1:250 000 scale and 'River' (feature code 5230) for 1:625 000 scale data. The use of different rivers and the discrepancies between the feature classifications at different scales will not affect this methodology, since, given that the feature being studied is modelled well by fractal geometry, the distance λ is a function of the data capture process, not the actual feature studied nor the feature classification. One of the data sets is shown inset within the Java application written for this analysis in Figure 2. It is noticeable that the lines do not form a continuous network due to the differential categorisation of different sizes of river. This lack of connectivity does not affect the methodology.

3.5 Experimentation

Computer code to calculate the number of dividers of varying lengths which describe the hydrological data was implemented in the Java object oriented programming language. Each line was tested by measuring $N(\delta)$, the number of dividers of length δ which describe the curve. The range of values of δ used started at the minimum mathematical precision of the current tile, ie the smallest distance that can possibly be represented by the data format (1cm in the case of 1:1250 and 1:2500 scale and 1m for 1:250 000 and 1:625 000 scale data). The values were increased by a factor of 2^n until $N(\delta)$ dropped to below 3.0, beyond which point further increases in δ tend to produce dividers comparable in length to or longer than the actual curve. One of the practical difficulties when calculating $N(\delta)$ is a consequence of δ rarely fitting an exact integer number of times round the feature (Carr and Benzer 1991). For this study, any remainder, in terms of the ratio of remaining curve to divider length, was simply added to $N(\delta)$, consequently $N(\delta)$ is not necessarily a whole number. The process was repeated for every line in the tile of length greater than 5% of the tile edge length. While this length is somewhat arbitrary, it was found to be about the minimum line length which allowed a suitably large range of scales to be tested. A small number of otherwise eligible lines had to be omitted from the analysis since they represented other linear features. The analysis was then repeated for each tile.

4 Results

Conventionally, the results of dividers analysis are presented in the form of a Richardson plot, showing $ln(\delta)$ against $ln(N(\delta))$ and using [4] below

$$D = \frac{ln(N(\delta))}{-ln(\delta)} \qquad \text{(Peitgen, Jurgens, and Saupe 1992)} \tag{4}$$

where D is the fractal dimension of the curve measured from the slope of the Richardson plot. However, the Richardson plot tends to obscure the predicted transition from fractal to linear behaviour, since the slope of the line and so the fractal dimension D would typically vary barely perceptibly from 1.0 for a non-fractal curve to 1.2 to 1.3 for most environmental data (Burrough 1981). Consequently, by taking advantage of the relationship:

$$L(\delta) = N(\delta).\delta \tag{5}$$

the results can be displayed as $ln(\delta)$ against the ratio of the natural log of observed length $L(\delta)$ to Euclidean metric length $d_E(x, y)$, shown in [6] below.

$$\frac{1-m}{ln(d_E(x,y))} = \frac{ln(L(\delta))/ln(d_E(x,y))}{ln(\delta)} \tag{6}$$

The results of the analysis for the four tiles studied are shown in Figure 3, using the ratio of observed to Euclidean metric lengths from Equation [6] as opposed to the Richardson plot which would be based on Equation [4]. For very small values of δ the ratio of observed to Euclidean metric length is approximately 1 for all the river sections, clearly shown in Figure 3. However, at a certain value of δ the graph quite abruptly changes as the observed length decreases sharply compared with the Euclidean metric length. The variety of slopes observed after this point are due to the slightly different fractal dimensions exhibited by the different curves. The distance at which the graph stops displaying a gradient of zero is the distance greater than which the representation of the river begins to display fractal behaviour. This distance is the resolution λ.

The ratio of the natural log of *total* observed length of river to *total* Euclidean metric length for each of the four scales is shown in Figure 4. It is of note that the curves in this diagram incorporate all of the data from each of the graphs in Figure 3 in a single data set respectively, although they are not simple averages of the data in Figure 3. Figure 4 uses the total observed length of sampled rivers at each scale, while Figure 3 takes no account of the different length of each river segment analysed. The four scales clearly show increasing distances on the x-axis at which the curve 'breaks' with decreasing scale.

4.1 Calculation of values for λ

The different values of λ interpreted from Figure 4 are shown in Table 1. The resolution λ is defined as the distance smaller than which the behaviour of the curve is linear and deterministic. For small values of δ the ratio of observed to Euclidean length is close to 1 whilst at a certain size of δ the observed length drops rapidly in comparison with the Euclidean length. From the definition above it follows that the resolution λ will be found where the ratio of observed to Euclidean length stops being close to 1. It is therefore possible to deduce the resolution λ from the curves in Figure 4 by setting a cut-off ratio of observed to Euclidean length above which the curve is behaving linearly and below which it is behaving fractally. Initially this cut off was arbitrarily defined as where the observed length became more than 0.1% lower than the Euclidean metric length, ie where the ratio of observed to Euclidean length drops below 0.9999. The values for λ obtained using this arbitrary cut-off are shown in Table 1.

Whilst this cut off has the advantage of simplicity it can be criticised since it has no statistical basis. For a better cut-off value we need to look more closely at the reasons why the ratio of observed to Euclidean length is not *precisely* 1 for small values of δ. One cause is simply a lack of mathematical precision; despite most of the calculations being performed using 32 bit floating point numbers, rounding errors will occur. Additionally, even for small values of δ the dividers method suffers from interference between the frequency of points in the the curve being analysed and the dividers length. The concept is illustrated by Figure 5; here a line of Euclidean length 10 can have an observed length of 10 for a dividers length of 1, but also have an observed length of slightly less than 10 for a smaller dividers length. In the latter case the dividers length does not coincide with the frequency of points in the line under analysis.

Bearing this in mind, an number of essentially straight lines similar to the example in Figure 5 were simulated with a range of angles at their apex. By measuring the ratio of observed to Euclidean length with a variety of small dividers lengths for this non-fractal simulated data, a sample data set describing the variability of observed to Euclidean length for non-fractal data was built up. This sample data set was used as the basis for deciding what cut-off ratio to use in calculating λ. Since the simulated lines were non-fractal, all the variability in the measurement of observed to Euclidean length can be attributed to rounding errors and interference. It follows that any observation which falls outside this sample distribution is significantly different from expected non-fractal behaviour. Consequently, the confidence interval of the sample distribution can be taken as the cut-off. The sample data set was comprised of 400 observations with a mean of 0.99989 and standard deviation of 0.00074. Under the assumption of normal distribution this translates into a 95% confidence interval of 0.00122 which in turn gives a cut-off of 0.99878. Table 1 details the values for the resolution λ derived from this cut-off ratio.

4.2 Minimum feature sizes

Table 1 also gives the minimum feature size for each map scale, assuming a level of cartographic error of 0.2mm. This magnitude of cartographic error is often taken to be the minimum level cartographic error attainable with skilled draughting (Maling 1989). Consequently, given that there are no features on a physical hardcopy map smaller than the cartographic error of 0.2mm, the corresponding feature size in the physical world can be deduced by multiplying the cartographic error by the denominator in the scale representative fraction, often termed the scale number (eg, minimum feature size for 1:1250 scale map is 0.2mm * 1250 = 0.25m). These minimum feature sizes are of use, since they express the expected resolution, for a paper map at a given scale, below which no features in the physical world are represented. The resolution λ expresses the same relation for digital maps, so the two measures are contrasted in the next section.

5 Discussion

The methodology presented here allows the assessment of a resolution λ for any curve stored as digital vector data, assuming that the feature represented by the curve is modelled well by fractal geometry. The distance λ itself represents the point at which the fractal model breaks down. However, the initial discussion of internal evidence in the introduction required not only that inconsistencies between the model and the data be identified, but also that these inconsistencies be attributed to data quality.

5.1 Interpretation of λ

Goodchild and Gopal (1989) lament the contradiction within GIS where the apparent mathematical precision of stored data can often far exceed the accuracy. The cause of this contradiction is that data storage in GIS can be to an arbitrarily high level of mathematical precision, unrelated to the resolution of the data capture equipment. The results suggest that the resolution λ is an artifact of the data capture method and equipment. By devising a method for actually calculating λ it is possible to address the contradiction posed above and set a limit on accuracy, since the resolution of the data capture equipment is important in determining the maximum possible information content of a data set, and it is axiomatic that the accuracy of a data set cannot exceed the information contained within that data.

Further, detail or mathematical precision is in itself an important factor in determining data quality. Whilst not part of the classic five elements of spatial data quality; lineage, positional accuracy, attribute accuracy, logical consistency and completeness (NCDCDS, 1988), increasingly such additional quality elements are seen as central to an exhaustive approach to quality. Goodchild and Proctor (1997) writes that

> “[the] level of geographic detail is a critical element in determining a data set’s fitness for a given use, it is important that effective methods be found for its characterisation.”

The assertion made here, then, is that the level of detail of positional data is an important element of data quality and as such the resolution λ expresses this level of detail for digital representations of fractal features.

5.2 Resolution and scale

The scale of digital geographic data is a problematic issue, as the ease with which digital data can be displayed without reference to any physical cartographic limitations means the conventional cartographic concept of scale is inappropriate. However, the assumption made here is that digital data does have an analogue of scale inasmuch as the level of detail in the data is dependent upon the finite resolution of the data capture methods and equipment. If the data capture resolution is known, can be assumed or calculated using the methodology presented here, some inferences can be made about the original map scale.

The results of such an assumption, where the λ values can be compared with the expected minimum resolution assuming a cartographic error of 0.2mm, are shown in Table 1. Keefer, Smith, and Gregoire (1991) have found cartographic errors of between 0.254 and 0.508mm to be reasonable, while as mentioned above other authors have made use of a cartographic error of 0.2mm. These results suggest quite strongly that 1:250 000 and 1:625 000 maps are the sources for the two smaller scale digital data sets, since the expected minimum resolution calculated in this way tallys well with the resolution λ calculated using both cut-off ratios. Here, the original map scale can be quite accurately deduced from λ and a simple assumption about cartographic error. However, the same cannot be said for the two digital data sets derived from the larger scale maps, since the expected minimum resolution is approximately 20 times smaller than that calculated using λ.

When considering this result it must be borne in mind that the underlying assumptions when attempting to relate λ to original map scale is not simply the magnitude of the cartographic error, but also that distance-based *stream mode digitising* is utilised and that no significant post digitising nor smoothing is performed on the data. Newby (1992) reports that large scale map digitising quality control procedures used by the OSGB require not only certain standards with regard to feature coding, positional accuracy, squareness of buildings and completeness but also prevent an excessive number of points being used in the representation of features. The quality control test on excessive points requires an expert (ie an experienced digitising staff member) to determine the minimum number of points to represent a feature and to count the number actually used. The difference should not exceed 25%. This quality control step presupposes *point mode digitising.* Practitioners have long advocated stream mode digitising for small scale maps and point mode digitising for large scale maps (Bell 1978; Bell and Bickmore 1978).

It is suggested that the satisfactory relation between λ and minimum feature size for digital data derived from small scale source maps is a reflection of stream mode data capture and that in such cases calculation of λ is a useful method of determining source map scale where it is not known. Conversely, the less satisfactory results for digital data derived from large scale maps may have arisen due to point mode digitising.

While the previous discussion indicates some limitations of the resolution λ in estimating the source map scale where it is unknown, λ can provide other lineage based quality information. Where source map scale is known, λ can provide a check upon such information and where discrepancies exist they can be attributed to data capture techniques as in the previous paragraph.

5.3 Methodology

The methodology presented here contrasts with that presented by Brunsdon and Openshaw (1994), outlined in §2.3, who made a point of producing an estimator of local line complexity, creating line segments homogeneous with respect to complexity. Their contention was that the location of the beginning and end of line segments contained within the raw data is geographically arbitrary. Indeed this contention was borne out during the course of this work, since much of the hydrological data contains breaks due to intersections with other features, such bridges. A counter argument would suggest that while the location of such breaks have limited geographic significance, they are likely to define line segments which are homogeneous with respect to quality, as an individual line segment will probably have been digitised at one time using one data capture method. In either event, the methodology presented here should be adaptable, indeed could provide a more quantitative second phase of a data quality assessment, following the classification of a feature representation into line segments using Brunsdon and Openshaw's method.

The main advantage of this methodology is that it does not depend on the measurement of fractal dimension, a process that is fraught and imprecise (Peitgen, Jurgens, and Saupe 1992), but instead takes advantage of the limitations of the fractal model as applied to digital environmental data. However, it is of note that the method is only intended as one tool amongst a variety which can be used to assess data quality and would be best employed within the context of wider data quality management. Where deductions can be made about the source map scale, this information could feed into further quality procedures such as the least squares adjustment of positional information discussed in §2.2.

5.4 Further work

The methodology set out in this paper offers the potential for a variety of avenues of exciting research. An attempt has been made here to relate λ to existing measures of data quality, such as lineage. The initial results are supportive of the concept, although further research into a wider variety of original map scales, particularly 1:10 000 and 1:50 000 and other digital map sources, such as Harper Collins (1:100 000) and Harvey (1:40 000) would need to be undertaken to verify the response of λ to the entire range of possible data capture methods.

The resolution λ reflects the minimum ground resolution of the map and as such may have some application as a measure of data quality in itself. We have already seen in §5.1 that a case can be made for inclusion of detail or mathematical precision within the important elements of data quality. Following further empirical testing of this analysis, it presents a method for calculating detail for fractal data sets directly.

With some algorithms the original resolution of raster data sets may survive the process of vectorisation. If this is the case, the method should also be applicable to digital data derived from raster sources to assess the resolution of the source data. Furthermore, by applying the method in the opposite direction it would be possible to perform quality control checks upon data capture where the minimum resolution of the data capture equipment is known. Here, the quality of the data capture might be assessed by comparison of the known resolution with λ.

While the example given here uses λ to assess data quality, λ is a sensitive index as to the level of detail in the representation of a fractal object in the physical world. Consequently, it may have considerable application in feature extraction and classification, allowing assessment of the representational characteristics of vector data on a per-feature basis. Finally, the analysis offers the potential for the assessment of the performance of generalisation routines in maintaining the characteristics of a fractal curve whilst reducing (or increasing) the data volume.

6 Conclusions

This paper has shown that for digital vector geographic data representing features modelled closely by fractal geometry, such as hydrology, there exists a distance λ, shorter than which the representation of the feature does not exhibit fractal characteristics. Using river features in OSGB digital data derived from 1:1250, 1:2500, 1:250 000 and 1:625 000 maps, this distance λ was calculated. The inference is made that the existence of non-fractal behaviour, characterised by differentiable length, is an artifact of the way digital vector data is represented in Euclidean coordinate space. Consequently the distance λ reflects the resolution of the equipment and methods used for data capture. Potentially, λ is of value in itself, since the level of detail may be an important element of data quality for positional data. However, using a limited number of assumptions it is also possible to make deductions about the source map scale. Calculation of λ for the river network data revealed a resolution very close to that expected, given the source map scale and assuming a cartographic error of 0.2mm for the small scale derived data. For the large scale derived data, λ was calculated to be much larger that would be expected. The deduction was made that the deliberate removal of excess points often associated with the point mode digitising data capture technique used for large scale maps may have produced a larger value of λ than expected, while stream mode digitising, usually used for smaller scale maps, would not be expected to produce such an effect.

References

Aalders, H. (1996). Quality metrics for GIS. In M.J. Kraak and M. Molenaar (Eds.), *Advances in GIS research II; Proceedings Seventh International Symposium on Spatial Data Handling*, Volume 1, pp. 5B1–10.

Atkinson, P. (1995). A method for describing quantitatively the information, redundancy and error in digital spatial data. In P. Fisher (Ed.), *Innovations in GIS 2*, pp. 85–96. Taylor and Francis: London.

Bannister, A., S. Raymond, and R. Baker (1992). *Surveying* (6th ed.)., pp. 239–274. Longman: Essex.

Bell, J. (1978). The development of the ordnance survey 1:250 000 scale derived digital map. *Cartographic Journal 15*(1), 7–12.

Bell, S. and D. Bickmore (1978). Interactive cartography at the ECU: Regional geography à la mode. In D. Merriam (Ed.), *Recent advances in geomathematics*, pp. 117–134. Pergamon Press: Oxford.

Brown, A. and C. van Elzakker (1993, May). The use of colour in the cartographic representation of information quality generated by a GIS. In *Proceedings 16th International Cartographic Conference*, Volume 2, Koln.

Brunsdon, C. and S. Openshaw (1993). Simulating the effects of error in GIS. In P. Mather (Ed.), *Geographical information handling: Research and applications.* John Wiley & Sons: Chichester.

Brunsdon, C. and S. Openshaw (1994). Error simulation in vector GIS using neural computing methods. In M. F. Worboys (Ed.), *Innovations in GIS 1*, Volume 1, Chapter 13, pp. 177–200. Taylor and Francis: London.

Burrough, P. (1981). Fractal dimension of landscapes and other environmental data. *Nature 294*, 240–242.

Carr, J. and W. Benzer (1991). On the practice of estimating fractal dimension. *Mathematical Geology 23*(7), 945–958.

Chrisman, N. (1991). The error component in spatial data. In D. Maguire, M. Goodchild, and D. Rhind (Eds.), *Geographical information systems*, Volume 1, Chapter 12, pp. 165–174. Longman: Essex.

Clarke, K. (1986). Computation of fractal dimension of topographic surfaces. *Computers and Geosciences 12*(5), 713–722.

Datcu, M. and K. Seidel (1994). Fractals and multi-resolution techniques for the understanding of geo-information. In G. Wilkinson, I. Kanellopoulos, and J. Megier (Eds.), *Fractals in geoscience and remote sensing*, Italy, pp. 56–84. Institute for Remote Sensing Applications.

David, B., M. van den Herrewegen, and F. Salgé (1996). Conceptual models for geometry and quality of geographic information. In P. Burrough and A. Frank (Eds.), *Geographic objects with indeterminate boundaries*, Volume 2 of *GIS Data*. Taylor and Francis: London.

Drummond, J. (1995). Positional accuracy. In S. Guptill and J. Morrison (Eds.), *Elements of spatial data quality*, Chapter 3, pp. 31–58. Elsevier Science: Oxford.

Drummond, J., A. Brown, D. Du, and C. van Elzakker (1996, May). The development of a GIS information quality model. In *Spatial accuracy assessment in natural resources and environmental sciences: 2nd International Symposium*, Fort Collins.

Feder, J. (1989). *Fractals*. New York: Plenum Publishing.

Goodchild, M. (1988). Lakes on fractal surfaces: a null hypothesis for lake rich landscapes. *Mathematical Geology 20*(6), 615–629.

Goodchild, M. and S. Gopal (1989). *Accuracy of spatial databases*, pp. xi–xv. Taylor and Francis: London. Preface.

Goodchild, M. and J. Proctor (1997). Scale in a digital geographic world. *Geographical and environmental modelling 1*(1), 5–23.

Goodchild, M. and N. Tate (1992). Description of terrain as a fractal surface and application to digital elevation model quality assessment. letter to photogrammetric engineering and remote sensing. *Photogrammetric Engineering and Remote Sensing 58*(11), 1568–1570.

Haines-Young, R. and J. Petch (1986). *Physical geography: its nature and methods*. Paul Chapman Publishing Ltd: London.

Harley, J. (1975). *Ordanance Survey Maps a descriptive manual*. Ordnance Survey: Southampton.

Heuvelink, G., P. Burrough, and A. Stein (1989). Propagation of errors in spatial modeling in GIS. *International Journal of Geographical Information Systems 3*(4), 303–322.

Howard, D. and A. MacEachran (1996). Interface design for geographic visualisation: Tools for representing reality. *Cartography and Geographic Information Systems 23*(2), 59–77.

Keefer, B., J. Smith, and T. Gregoire (1991). Modelling and evaluating the effects of stream mode digitising errors on map variables. *Photogrammetric Engineering and Remote Sensing 57*(7), 957–963.

La Barbera, P. and R. Rosso (1989). On the fractal dimension of river networks. *Water Resources Research 25*(4), 735–741.

Maling, D. (1989). *Measurement from maps*. Pergammon Press: Oxford.

Mandelbrot, B. (1982). *The fractal geometry of nature*. W.H.Freeman and Co: New York.

Mikhail, E. (1978). *Observations and least squares*. IEP Dun Donnely: New York.

Monckton, C. (1994). An investigation into the spatial structure of error in digital elevation data. In M. F. Worboys (Ed.), *Innovations in GIS*, Volume 1, Chapter 1, pp. 201–215. Taylor and Francis: London.

National Committee for Digital Cartographic Data Standards (1988). The proposed stran-dard for digital catographic data. *American Cartographer 15*(1), 11–142.

Newby, P. R. T. (1992). Quality management for surveying, photogrammetry and digital mapping at the Ordnance Survey. *Photogrammetric Record 14*(79), 45–58.

Openshaw, S., M. Charlton, and S. Carver (1991). Error propagation: A Monte Carlo simulation. In I. Masser and M. Blakemore (Eds.), *Handling geographical information*, pp. 78–101. Longman: New York.

Ordnance Survey of Great Britain (1997). Data collection resource study modelling exercise. Restricted circulation.

Peitgen, H., H. Jurgens, and D. Saupe (1992). *Chaos and fractals: New frontiers of science*. New York: Springer-Verlag.

Pentland, A. (1984). Fractal-based description of natural scenes. *IEEE transactions on pattern analysis and machine intelligence PAMI-6*(6), 661–674.

Phillips, J. (1993). Interpreting the fractal dimension of river networks. In N.-N. Lam and L. De Cola (Eds.), *Fractals in geography*, Chapter 7, pp. 142–157. Prentice Hall: New Jersey.

Pohl, C. (1996). Geometric aspects of multisensor image fusion for topographic map updating in the humid tropics. In *ITC Publication 39*. ITC: Netherlands.

Polidori, L. (1994). Fractal-based evaluation of relief mapping techniques. In G. Wilkinson, I. Kanellopoulos, and J. Megier (Eds.), *Fractals in geoscience and remote sensing*, Italy, pp. 277–297. Institute for Remote Sensing Applications.

Polidori, L., J. Chorowicz, and R. Guillande (1991). Description of terrain as a fractal surface and application to digital elevation model quality assessment. *Photogrammetric Engineering and Remote Sensing 57*(10), 1329–1332.

Raper, J. and D. Livingston (1995). Development of a geomorphological spatial model using object oriented design. *International Journal of Geographical Information Systems 9*(4), 359–384.

Robert, A. and A. Roy (1990). On the fractal interpretation of the mainstream length-drainage area relationship. *Water Resources Research 26*(5), 839–842.

Spear, M., J. Hall, and R. Wandsworth (1996, May). Communication of uncertainty in spatial data to policy makers. In *Spatial accuracy assessment in natural resources and environmental sciences: 2nd International Symposium*, Fort Collins, pp. 199–207.

Stanislawski, L., B. Dewitt, and R. Surestha (1996). Estimating positional accuracy of data layers within as GIS through error propagation. *Photogrammetric Engineering and Remote Sensing 62*(4), 429–433.

Tveite, H. and S. Langaas (1999). An accuracy assessment method for geographical line data sets based on buffering. *International Journal of Geographical Information Science 13*(1), 27–47.

Unwin, D. (1995). Geographical information systems and the problem of error and uncertainty. *Progress in Human Geography 19*(4), 549–558.

van Elzakker, C., B. Ramlal, and J. Drummond (1992). The visualisation of GIS generated information quality. In *Archives ISPRS Congress XVII*, Volume 29.B4, pp. 608–615.

Veregin, H. (1995). Developing and testing an error propagation model for GIS overlay. *International Journal of Geographical Information Systems 9*(6), 595–616.

Wood, J. (1996). *The geomorphological characterisation of digital elevation models*. Ph. D. thesis, Department of Geography, University of Leicester.

Xia, Z.-G. and K. Clarke (1997). Approaches to scaling of geo-spatial data. In D. Quattrochi and M. Goodchild (Eds.), *Scale in remote sensing and GIS*, Chapter 15, pp. 309–360. CRC Press: New York.

Scale	1:1250	1:2500	1:250 000	1:625 000
Results for cut-off of 0.1% from Figure 4				
Dividers length $ln(\delta)$	6.245	7.209	8.757	9.445
Resolution λ (m)	5	14	64	126
Results for 95% confidence interval from Figure 4				
Dividers length $ln(\delta)$	6.353	7.378	8.997	9.627
Resolution λ (m)	6	15	81	152
Minimum feature size (m) (using cartographic error of 0.2mm)	0.25	0.5	50	125

Table 1: Resolution λ values and minimum feature sizes

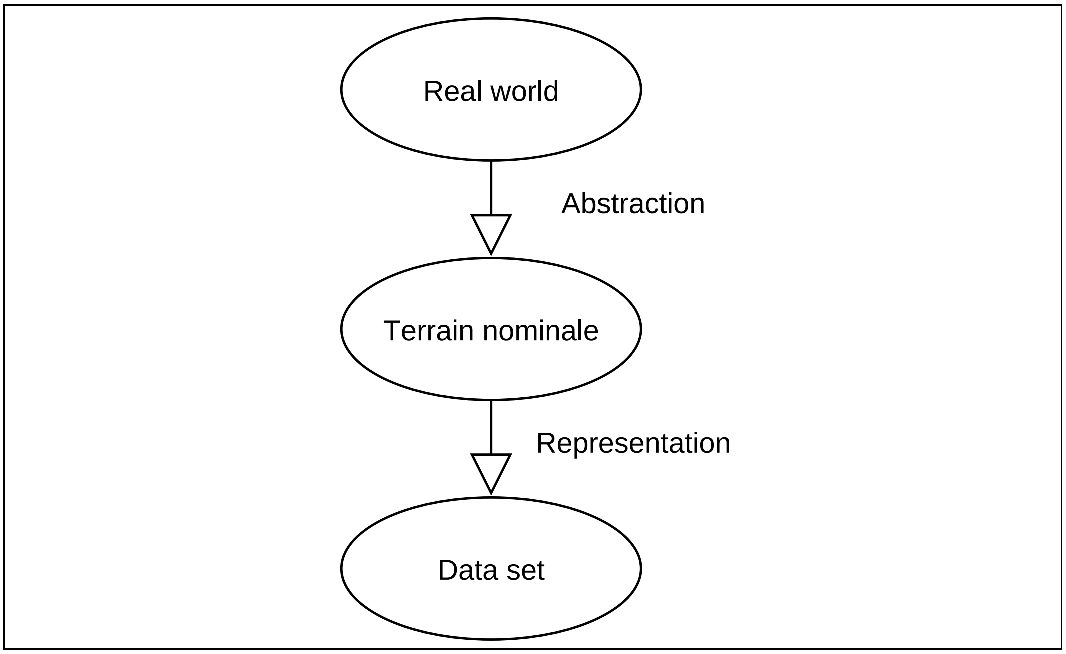
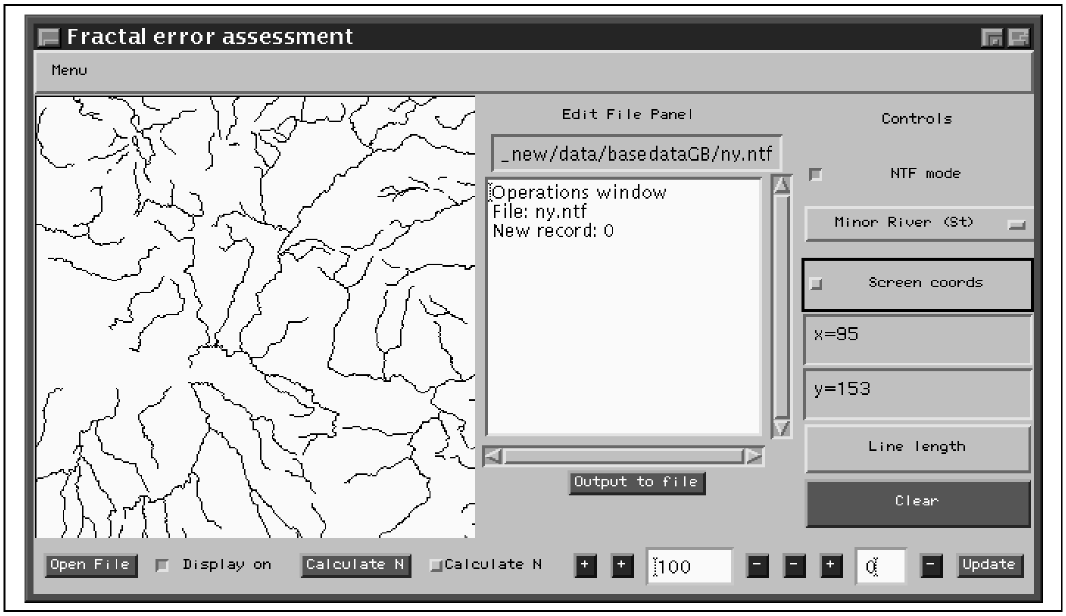


Figure 1: Construction of terrain nominale



_new/data/basedataGB/ny.ntf

```
Operations window
File: ny.ntf
New record: 0
```

Minor River (St)

x=95

y=153

100

0

Figure 2: Java applet with sample data display inset
Hydrological data reproduced with the permission of the Controller of Her Majesty's Stationary office.
©Crown Copyright Licence no. ED 274259

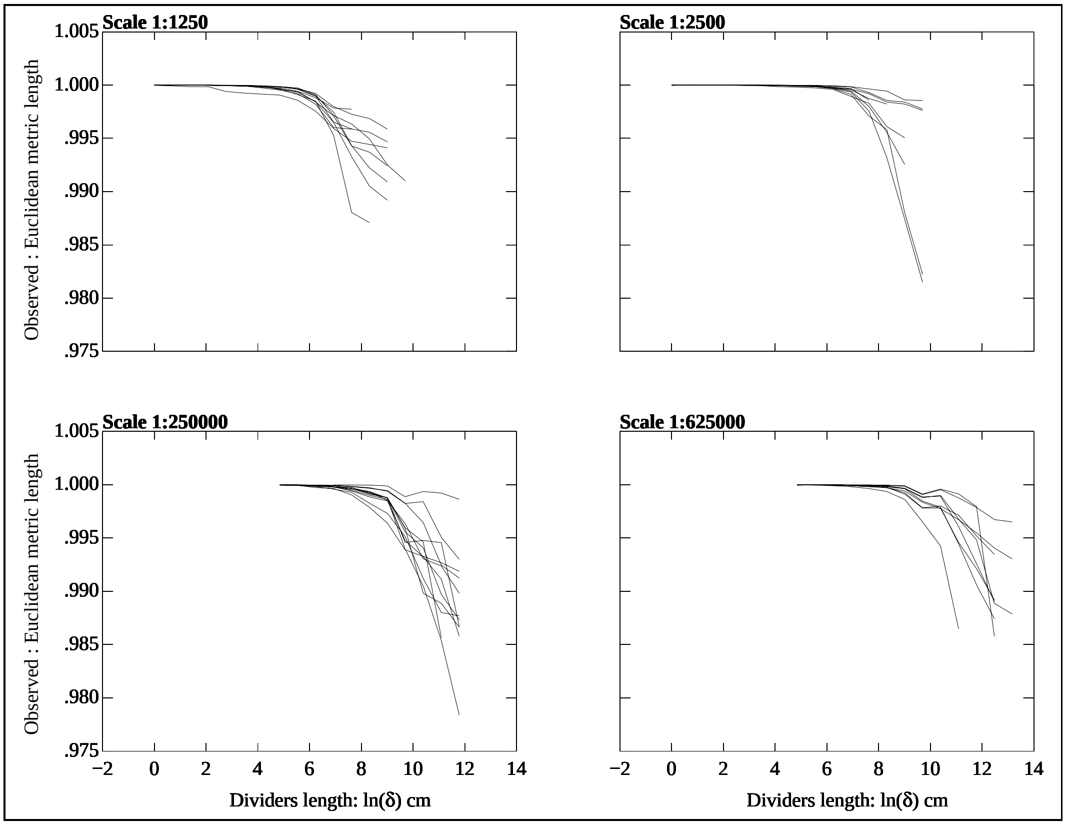


Figure 3: Changes in observed river lengths for a variety of river segments at four map scales

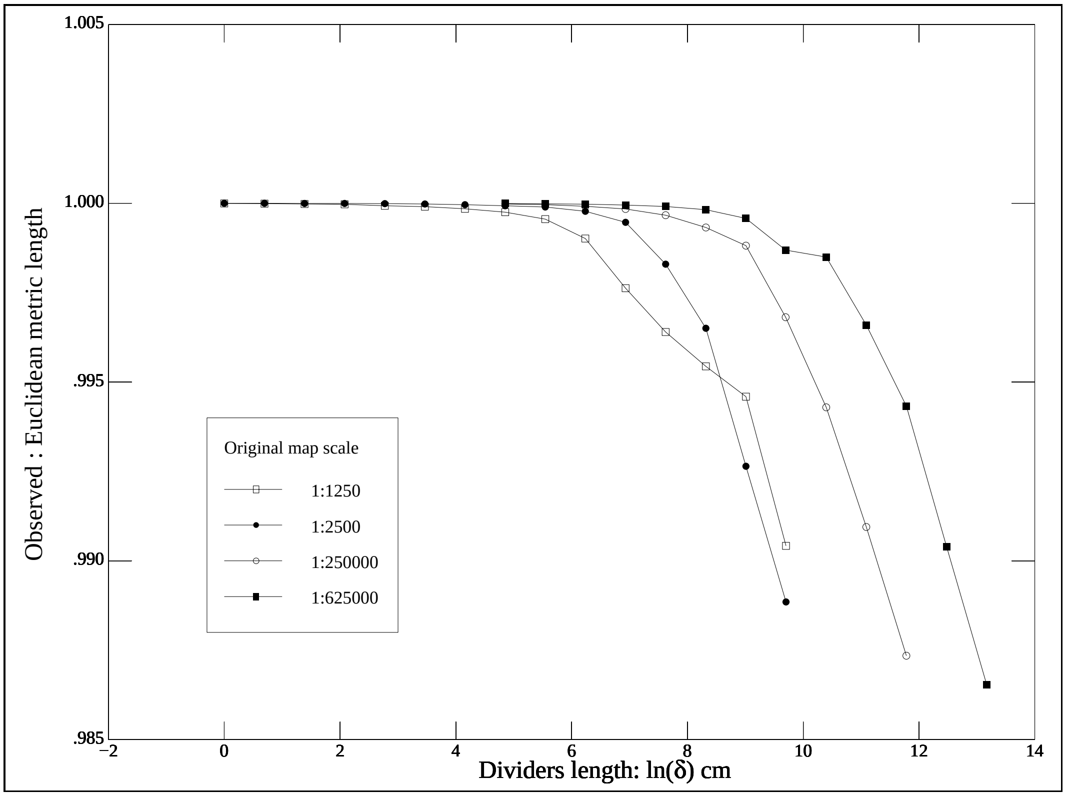


Figure 4: Changes in total river network length at four map scales

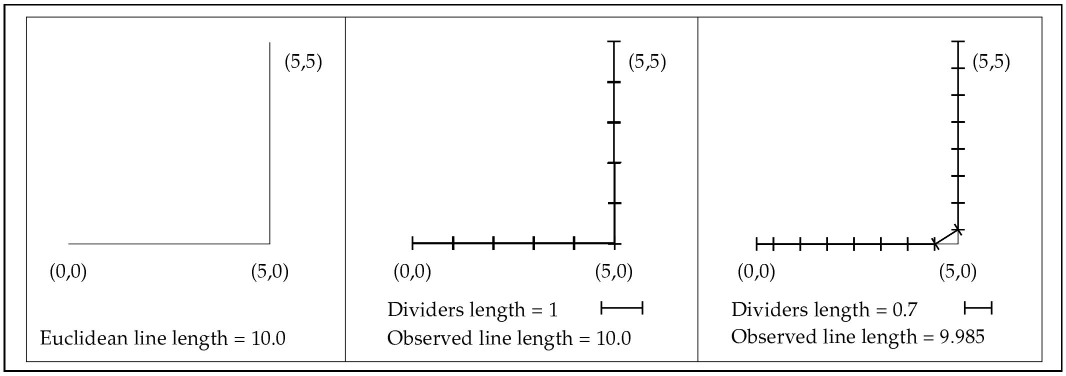


Figure 5: Effects of interference between line and dividers length

University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
DUCKHAM, MATT;Drummond, Jane

Title:
Assessment of error in digital vector data using fractal geometry

Date:
2000

Citation:
Duckham, M., & Drummond, J. (2000). Assessment of error in digital vector data using fractal geometry. International Journal of Geographical Information Science, 14(1), 67-84.

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/34960